TU WIEN Mathematics

Scrambling permutations and related structures

Asymptotics and Constructions

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Technical Mathematics

by

Enrico Iurlano, BSc
Registration Number 01427258

to the Institute of Discrete Mathematics and Geometry

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Bernhard Gittenberger

Vienna, 12th September, 2022

Enrico Iurlano	Bernhard Gittenberger

TU WIEN Mathematics

Scrambling Permutationen und verwandte Strukturen

Asymptotik and Konstruktionen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Technische Mathematik

eingereicht von

Enrico Iurlano, BSc
Matrikelnummer 01427258

am Institut für Diskrete Mathematik und Geometrie

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Bernhard Gittenberger

Wien, 12. September 2022

Enrico Iurlano	Bernhard Gittenberger

Erklärung zur Verfassung der Arbeit

Enrico Iurlano, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. September 2022

Enrico Iurlano

Acknowledgements

First and most of all, I would like to thank my supervisor Bernhard Gittenberger for very helpful discussions during the preparation of this thesis. He was formative for me through his lectures/exercises in the field of discrete mathematics and, apart from this thesis, has kindly given me further hints/comments concerning a related paper submission.

My fellow students deserve special thanks for the countless hours and nights spent together solving problems and for the good time and shared interests even away from the subject matter.

I would like to thank my parents for moral and financial support, which was very helpful during my studies. A great thank you goes to my girlfriend for many honest opinions and precious advice.

Abstract

A family of permutations of the symmetric group S_n is called k-scrambling if for any sequence of pairwise distinct positions $(p_1, \ldots, p_k)$ there exists a permutation whose evaluation in p_1 is minimal among the remaining evaluations in p_j. Moreover, the more special property of being completely k-scrambling requires that the successive evaluations in $p_1, \ldots, p_k$ form a monotonically increasing sequence. Families satisfying additional sharpenings of the latter two properties are k-restricted min-wise independent or, in the second, stronger case coincide (up to isomorphic identification) with perfect sequence covering arrays (PSCAs) of strength k. Determining permutation families that satisfy these differently strong properties and have small (or even minimal possible) cardinality is non-trivial and of great importance for other mathematical branches (order theory, geometry, graph theory, etc.) as well as for a variety of practical applications such as automated testing of software and fast similarity estimation of documents for web search engines. We analyze the asymptotic behavior of these smallest possible cardinalities and also address algorithmic approaches to construct such families explicitly. Here we succeed in answering positively an open question of R. Yuster on the polynomial boundedness of the cardinalities of PSCAs [Yus20]; we also obtain improved asymptotic lower bounds and for $k = 3$ and $k = 4$ improved asymptotic upper bounds. Even in the non-asymptotic range – i.e., for concrete, small values of n – we achieve some improvements for $k = 3$ and $k = 4$ (up to a factor of 7.5 compared to the currently available state of the art in the literature). Finally, we propose a new class of permutation families that satisfy a certain reflection symmetry, and use it to restrict the search space that needs to be scanned in order to computationally seek PSCAs systematically via backtracking. This class could potentially be useful to achieve further improvements in the non-asymptotic range in the future with sufficient computational power.

Kurzfassung

Eine Familie von Permutationen der symmetrischen Gruppe S_n heißt k-scrambling, falls zu jeder beliebigen Abfolge paarweise verschiedenener Positionen $(p_1, \ldots, p_k)$ eine Permutation existiert, deren Auswertung in p_1 minimal unter den restlichen Auswertungen in p_j ist. Die speziellere Eigenschaft vollständig k-scrambling zu sein, erfordert darüber hinaus, dass die sukzessiven Auswertungen in $p_1, \ldots, p_k$ eine monoton steigende Folge bilden. Familien, die zusätzlichen Verschärfungen letzterer beider Eigenschaften genügen, nennt man k-restricted min-wise independent bzw. stimmen im zweiteren, stärkeren Fall (bis auf isomorphe Identifikation) mit Perfect Sequence Covering Arrays (PSCAs) der Stärke k überein. Die Bestimmung von Permutationsfamilien, die diesen unterschiedlich starken Eigenschaften genügen und möglichst kleine (oder gar minimal mögliche) Kardinalität besitzen, ist nicht-trivial und von großer Bedeutung für andere mathematische Richtungen (Ordnungstheorie, Geometrie, Graphentheorie etc.) sowie für eine Vielzahl praktischer Anwendungen wie etwa automatisiertes Testing von Software und schnelle Ähnlichkeitsschätzung von Dokumenten für Suchmaschinen. Wir analysieren das asymptotische Verhalten dieser kleinstmöglichen Kardinalitäten und befassen uns auch mit algorithmischen Zugängen, um solche Familien explizit zu konstruieren. Hierbei gelingt es uns eine offene Frage von R. Yuster nach der polynomialen Beschränktheit der Größen von PSCAs [Yus20] positiv zu beantworten; wir erhalten außerdem verbesserte asymptotische untere Schranken sowie für $k = 3$ und $k = 4$ verbesserte asymptotische obere Schranken. Auch im nicht-asymptotischen Bereich – also für konkrete, kleine Werte von n – erreichen wir für $k = 3$ und $k = 4$ einige Verbesserungen (bis zu Faktor 7.5 im Vergleich zum derzeit verfügbaren State of the art in der Literatur). Abschließend schlagen wir eine neue Klasse von Permutationsfamilien vor, die einer gewissen Spiegelungssymmetrie genügen, und nutzen sie zur Ausdünnung des Suchraumes, der überprüft werden muss, um PSCAs computergestützt systematisch via Backtracking zu suchen. Diese Klasse könnte womöglich nützlich sein, um mit ausreichend Rechenleistung künftig weitere Verbesserungen im nicht-asymptotischen Bereich zu erzielen.

Contents

Abstract **ix**

Kurzfassung **xi**

Contents **xiii**

1 Introduction **1**
1.1 Historical development and motivation 1
1.2 Goals and contributions of this thesis 2
1.3 Thesis structure 3

2 Basic concepts **5**
2.1 Notation 5
2.2 Terminology and concepts 6
2.3 Related combinatorial structures 11

3 Methods **19**
3.1 The probabilistic method 19
3.2 Information theory for extremal combinatorics 20
3.3 Deletion correcting codes 23

4 Asymptotics **27**
4.1 k-scrambling permutations 27
4.2 Completely k-scrambling permutations 31
4.3 k-restricted min-wise independent permutations 42
4.4 PSCAs and rankwise independent permutations 45
4.5 Comparison of bounds 50

5 Constructions of completely scrambling families **51**
5.1 Deterministic polynomial time construction keeping asymptotic bounds 51
5.2 Tarui's construction for strength three 55
5.3 Generation of k-scrambling permutations: further approaches 57

6 Constructions of Perfect Sequence Covering Arrays **59**

6.1 Construction via Varshamov-Tenengolts codes 59
6.2 Computational constructions for PSCAs 61
6.3 Feasible embeddings of m-sequences 62
6.4 General search with pre-determined distributions 69

7 Applications **73**
7.1 Order theory, combinatorial geometry and related areas 73
7.2 Combinatorial software testing 74
7.3 Min-wise hashing 76

8 Conclusion **79**
8.1 Discussion 79
8.2 Open questions and further work 80

List of Figures **83**

List of Tables **85**

List of Algorithms **87**

Bibliography **89**

CHAPTER 1

Introduction

1.1 Historical development and motivation

The property of k-scramblingness was originally introduced under the naming "k-suitability" by B. Dushnik in the 1950s shortly after he had come up with the notion of the dimension of a partial order [DM41]. Twenty years later, a work of J. Spencer, [Spe71], setting an emphasis on the minimum size of such families, attributed interest on its own to the concept. It seems that the interest on how small such families can be kept was initially driven by the minimality sought-after in the context of the dimension of partial orders. Later, seeking small upper bounds for such families became an independent central research question, especially for fixed, small values of k. In addition, Spencer highlighted a special case, which he called *completely* k-scrambling permutations [Spe71], that subsequently became of major interest and which still leaves, as we will see, some unresolved questions. Spencer's view on the structure is an order theoretic one: Completely scrambling families represent large enough collections of linear orders on $\{1, \ldots, n\}$, such that each subset of k elements $\{i_1, \ldots, i_k\}$ forms an ascending chain with respect to at least one of the linear orders – the family *scrambles* an arbitrary order on the set $\{i_1, \ldots, i_k\}$ in all possible $k!$ ways.

In the 1990s, during the exponential growth of the World Wide Web [BCFM00] and the need to efficiently determine the similarity of a large number of documents, the class of *k-restricted min-wise independent families* was introduced by Broder et al. in [BCFM00] (in their most general form such families follow an arbitrary probability distribution). It turns out to be a special case of scrambling permutations under specific circumstances.

A class being isomorphic to the class of completely k-scrambling families was independently proposed for the purpose of event sequence testing in [KHL+12] and termed *sequence covering array of strength k*. It considers permutations as strings of (distinct) characters in $\{1, \ldots, n\}$ which collectively ensure to host (as subsequence) any string of length k

(composed of distinct symbols). This has subsequently launched a renaissance of interest in the base idea of (complete) k-scramblingness. In particular, a focus on the explicit computation of such families (of reduced/minimum size) has crystallized, and when the enormous computational effort even for small n was noticed (cf. [BTI12, BEI+12]), the interest shifted to the search for efficient approximation algorithms [KHL+12, BEI+12, CCHZ13], as well as to the further development of constructions that were initially conceived for the estimation of asymptotic bounds of these minimal sizes of families (cf. [CCHZ13, Tar08]).

Finally, the aforementioned renaissance also led to the study of *directed designs* from the viewpoint of SCAs possessing the property of *perfectness* (which enforce a much more regular and particular structure) [Yus20]. The naming seems to be inspired by a connection to *perfect codes* from coding theory [Lev91]. Directed designs, which are objects residing in the area of combinatorial design theory, although having been studied earlier (see e.g. [Lev91, MvT99]), seemed however not to have been examined in full generality and not with great attention for asymptotic growth. This question is addressed in the work [Yus20], which besides has asked for and successfully launched some interesting developments to the explicit construction of PSCAs [Na21, GW22, NJL22]. The work [Yus20] comes up with asymptotic polynomial lower bounds and asks about the existence of polynomial upper bounds for the size of PSCAs; so far only super-exponential (trivial) bounds are known.

Scrambling permutations (and derivatives) are a good example of a combinatorial structure whose development has benefited significantly from a constant interplay between abstract exploration of the structure and the use/modification of the structure to solve practical problems.

1.2 Goals and contributions of this thesis

Goals. Firstly, the manifold approaches in the literature (often dealing with isomorphic or similar concepts) to the topic are to be presented in an as far as possible harmonized manner. The relevance for further research areas and especially for applications shall also be made clear. One main focus is to give a good insight into the state of the art of available asymptotic bounds. Greater attention is dedicated to the analysis, comparison, evaluation, and potential for optimization of these bounds. Recent developments especially concerning constructive and computer-aided approaches shall be discussed as well.

Contributions. We come up with a detailed comparison of the many structures arisen in the context of scrambling permutations. As a result, we obtain the insight that the theory of combinatorial (directed) designs and the younger theory concerning min-wise independent families can successfully be combined – this seems to be unnoticed so far, and leads to the settlement of polynomial boundedness for PSCAs, which was an open question posed by R. Yuster in 2020 [Yus20]. Moreover, we show that bounds for PSCAs/rankwise independent and min-wise independent families can be slightly improved. Recent developments concerning computational constructions of PSCAs [Na21, NJL22]

are enriched by the proposal to incorporate an additional symmetry requirement into the search of PSCAs.

1.3 Thesis structure

Chapter 2 introduces and illustrates the main concepts, and clarifies some relationships between them. Chapter 3 collects results from information and coding theory, which will be needed to prove asymptotic or constructive results. Chapter 4 concerns the asymptotic behavior of the minimum possible cardinality of k-scrambling families: It gradually proceeds from the most general form of k-scramblingness to the most specific. In particular, the chapter contains the answer to R. Yuster's question on the boundedness of the size of PSCAs together with further optimizations. In Chapter 5 the focus is set on explaining algorithmic constructions of completely k-scrambling families and on giving an overview over recently pursued approaches. Chapter 6 deals with the problem of constructing PSCAs; it explains the connection to a branch of coding theory, addresses the discussion of very recent computer-aided approaches/heuristics and illustrates the proposal to search for PSCAs within a new class of permutation families satisfying a particular symmetry property. Fields of application connected to acquisitions for scrambling permutations and derivatives are to be illustrated in Chapter 7. Lastly, the conclusion in Chapter 8 highlights the insights gained and provides an outlook on questions arising from them, which seem interesting for further research.

CHAPTER 2

Basic concepts

2.1 Notation

We denote the set of integers as $\mathbb{Z}$ and the non-negative integers (respectively positive integers) as $\mathbb{N}$ (respectively $\mathbb{N}^\times$). We use the abbreviation $[n] := \{1, \ldots, n\}$. For a set A, by $|A|$ we refer to its cardinality and by 2^A to its power set. The indicator function of A is denoted by $\mathbb{1}_A$. If A is a finite subset of $\mathbb{Z}$ and contains the element a, the rank of a in A is defined as $\mathrm{rank}(a, A) := |\{b \in A : b \leq a\}|$. We use the abbreviation *(unordered) k-set* to refer to a set of cardinality $k \in \mathbb{N}$. Analogously, we use the expressions *(ordered) k-tuple*, or equivalently *ordered k-set*, to speak about tuples consisting of precisely k entries.

It will be convenient to denote permutations in tuple notation, i.e., given a permutation $\pi : [n] \to [n]$, enlist the ordered evaluations in a tuple $(\pi(1), \ldots, \pi(n)) \in [n]^n$. Speaking set theoretically these tuples are precisely the bijective functions on $[n]$, i.e., members of the symmetric group S_n. As there are further notations of permutations (such as the cyclic notation) we emphasize that throughout the thesis we consistently use the aforementioned tuple notation. The permutation $(1, \ldots, n) \in S_n$ is abbreviated as id.

We define the set

$$S_{n,k} := \Big\{(a_1, \ldots, a_k) \in [n]^k : i \neq j \Rightarrow a_i \neq a_j\Big\} \subseteq [n]^k, \tag{2.1}$$

as the set of all k-tuples consisting of pairwise distinct entries.

We write $\binom{[n]}{k}$ for the set of all k-subsets of $[n]$; its cardinality is determined by the binomial coefficient $\binom{n}{k}$. We denote by $n!$ the factorial and by

$$!n := n! \sum_{j=0}^{n} (-1)^j / j! \tag{2.2}$$

the *subfactorial* of n.

For a group G and a subgroup H of G we express this relation by writing $H \leq G$. For $g \in G$, by $\langle g \rangle$ we mean the subgroup of G generated by g. If a is a divisor of b we write $a|b$. The greatest common divisor (respectively least common multiple) of a set of natural numbers M is denoted as $\gcd M$ (respectively $\operatorname{lcm} M$). For a real number x, $\lceil x \rceil$ (respectively $\lceil x \rceil$) determine the numbers $\min\{z \in \mathbb{Z} : x \leq z\}$ (respectively $\max\{z \in \mathbb{Z} : z \leq x\}$).

We use the Landau notation $f(n) = O(g(n))$ for expressing boundedness of f by g up to a constant factor (for sufficiently large n). Similarly, $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$, and $f(n) = \Theta(g(n))$ if $f(n) = O(g(n)) \wedge f(n) = \Omega(g(n))$. If f asymptotically dominates g, we denote this relation by $f(n) = o(g(n))$.

Our main objects will be families indexed by $[d]$ ($d \in \mathbb{N}^\times$) containing d not necessarily distinct permutations (abbreviated as d-family). For convenience, when displaying explicitly such a d-family $\mathscr{F}$, we often tacitly print/identify it as $d \times n$ matrix such that the i-th row contains at column index j the entry $\pi_i(j)$, where π_i is the i-th permutation of $\mathscr{F}$, i.e., $\mathscr{F}$ is printed as

$$\begin{bmatrix} \pi_1(1), & \ldots, & \pi_1(n) \\ \vdots & \ldots, & \vdots \\ \pi_d(1), & \ldots, & \pi_d(n) \end{bmatrix} \in [n]^{d \times n}.$$

In this context, we use A^T to denote the transposed of a matrix A.

2.2 Terminology and concepts

In the following we review certain collections of permutations satisfying a variety of special constraints regarding monotony specifications (they are illustrated by concrete instances in the subsequent Example 2.2.4). Some authors (cf. [BCG+16]) say that these families have certain *separation* properties.

Definition 2.2.1 ([Dus50, Für96])**.** *A family $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$ is said to be* k-scrambling *($k \in [n]$) if for any tuple $(p_1, \ldots, p_k) \in S_{n,k}$, there is $\pi \in \mathscr{P}$ such that $\pi(p_1) < \pi(p_j)$, for $j = 2, \ldots, k$. The minimum cardinality d, for which a k-scrambling d-family of permutations of $[n]$ exists, is denoted as $N(n, k)$.*

To check if a family is k-scrambling, we can check if for every unordered k-set $q = \{q_1, \ldots, q_k\} \in \binom{[n]}{k}$ of "positions" the following condition is fulfilled: After picking a distinguished element q^* from q, there is a permutation of the family which, evaluated in q^*, is minimal among all evaluations of the permutation in the positions in q. The previous property can be strengthened as follows.

Definition 2.2.2 ([Spe71, Für96])**.** *A family $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$ is called* completely k-scrambling *($k \in [n]$) if for every k-tuple $(p_1, \ldots, p_k) \in S_{n,k}$, there exists an element $\pi \in \mathscr{P}$ such that $(\pi(p_j))_{j=1}^k$ forms an ascending chain, i.e. $\pi(p_1) < \pi(p_2) < \ldots < \pi(p_k)$.*

The minimum cardinality d, for which a completely k-scrambling d-family of permutations of $[n]$ exists, is denoted as $N^*(n,k)$.

In view of the last definition, it will be convenient to denote for a position selection $p = (p_1, \ldots, p_k) \in S_{n,k}$, by $\mathcal{ASC}_p$ the set of all permutations $\pi \in S_n$ such that $\pi(p_1) < \pi(p_2) < \ldots < \pi(p_k)$. If the value of k is needed to be passed as additional information, we write $\mathcal{ASC}_p^{[k]}$. Moreover let us say that permutations in $\mathcal{ASC}_p$ are *ascending along* p.

In [Für96], Füredi analyzes the property of being 3-*mixing*. We state it in a more general form and call it the property of being k-mixing ($k \geq 3$). Without attributing a name to this concept appears in a proof in [Rad03]. In this thesis it is used exclusively as auxiliary property.

Definition 2.2.3. *For $n \geq k \geq 3$, we say that a family $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$ is* k-mixing *if for every k-tuple $(p_1, p_2, \ldots, p_{k-2}, p_{k-1}, p_k) \in S_{n,k}$, there exists a permutation $\pi \in \mathscr{P}$ such that $\pi(p_2) < \pi(p_3) < \ldots < \pi(p_{k-2}) < \pi(p_{k-1})$ and, moreover, either $\pi(p_1) < \pi(p_2) \wedge \pi(p_{k-1}) < \pi(p_k)$ or $\pi(p_k) < \pi(p_2) \wedge \pi(p_{k-1}) < \pi(p_1)$. We denote by $N^{\mathrm{mix}}(n,k)$ the smallest cardinality a family of k-mixing permutations of $[n]$ can possess.*

As in [CCHZ13] let us term the parameter k specifying completely k-scrambling permutations as *strength*, and let us use this notion also in the less specific setting of scrambling and mixing families of permutations. Moreover, let us say that families having minimal possible cardinality are *optimal*.

In the following we provide illustrative instances of permutation families, which demonstrate how laborious it is to verify the aforementioned properties already for $n = 4$ and $k = 3$.

Example 2.2.4. *Let $n = 4$, $k = 3$ and consider the following families of permutations:*

$$\mathscr{O} = \begin{bmatrix} 1 & 3 & 2 & 4 \\ 3 & 1 & 2 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix} \quad \mathscr{P} = \begin{bmatrix} 2 & 1 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 4 & 2 & 3 & 1 \\ 4 & 3 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \quad \mathscr{Q} = \begin{bmatrix} 2 & 3 & 1 & 4 \\ 1 & 4 & 2 & 3 \\ 4 & 1 & 2 & 3 \\ 2 & 3 & 4 & 1 \\ 2 & 1 & 4 & 3 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$$

The family $\mathscr{O}$ is an instance of a 3-scrambling 3-family of permutations (for $n = 4$, the smallest possible cardinality still permitting 3-scramblingness is indeed $d = 3$, cf. Proposition 2.2.6). We can see that all monotonicity specifications are satisfied: For each $(p_1, p_2, p_3) \in S_{4,3}$ with $p_1 \in \{1, 2, 4\}$, there exists $\pi \in \mathscr{O}$ attaining the value 1 at position p_1. If $p_1 = 3$, then $\pi(p_1) = 2$ for all $\pi \in \mathscr{O}$, and for each set of positions

$\{p_2, p_3\} \in \binom{\{1,2,4\}}{2}$ *we find* $\pi \in \mathscr{O}$ *such that* $\{\pi(p_2), \pi(p_3)\} = \{3, 4\}$ *being minorized by* $\pi(p_1)$.

We notice that $\mathscr{P}$ *is* 3*-mixing as all monotonicity requirements are met (cf. Table 2.1). Moreover, by examining every family consisting of just* 4 *permutations in* S_4 *one consistently obtains the violation of at least one condition in Definition 2.2.3 and, as a consequence, for* $n = 4$, $\mathscr{P}$ *has minimal cardinality among all families being* 3*-mixing.*

The last collection $\mathscr{Q}$ *is completely* 3*-scrambling (all monotonicity constraints are satisfied, as we show in Table 2.2) and* $|\mathscr{Q}| = N^*(4, 3) = 3!$*, i.e.,* $\mathscr{Q}$ *is of minimal possible cardinality (cf. Proposition 2.2.6).*

$(p_1, p_2, p_3) = (a_1, *, a_2)$		$\{i : \pi_i \in ASC_{(p_1,p_2,p_3)} \cup ASC_{(p_3,p_2,p_1)}\}$
$(a_1, 1, a_2)$,	$(a_1, a_2) = (2, 3)$	$\to \{1, 2\}$
	$(a_1, a_2) = (2, 4)$	$\to \{1\}$
	$(a_1, a_2) = (3, 4)$	$\to \{2\}$
$(a_1, 2, a_2)$,	$(a_1, a_2) = (1, 3)$	$\to \{4, 5\}$
	$(a_1, a_2) = (1, 4)$	$\to \{3, 4, 5\}$
	$(a_1, a_2) = (3, 4)$	$\to \{3\}$
$(a_1, 3, a_2)$,	$(a_1, a_2) = (1, 2)$	$\to \{3\}$
	$(a_1, a_2) = (1, 4)$	$\to \{1, 3, 5\}$
	$(a_1, a_2) = (2, 4)$	$\to \{1, 5\}$
$(a_1, 4, a_2)$,	$(a_1, a_2) = (1, 2)$	$\to \{2\}$
	$(a_1, a_2) = (1, 3)$	$\to \{4\}$
	$(a_1, a_2) = (2, 3)$	$\to \{2, 4\}$

Table 2.1: Explicit verification of the property of being 3-mixing for $\mathscr{P} = (\pi_1, \ldots, \pi_5)$ of Example 2.2.4.

(p_1, p_2, p_3)	$\{i : \pi_i \in \mathcal{ASC}_{(p_1,p_2,p_3)}\}$
$(1, 2, 3)$	$\{4\}$
$(1, 2, 4)$	$\{1\}$
$(1, 3, 2)$	$\{2\}$
$(1, 3, 4)$	$\{2\}$
$(1, 4, 2)$	$\{2\}$
$(1, 4, 3)$	$\{5\}$
$(2, 1, 3)$	$\{5\}$
$(2, 1, 4)$	$\{5\}$
$(2, 3, 1)$	$\{3\}$
$(2, 3, 4)$	$\{3\}$
$(2, 4, 1)$	$\{3\}$
$(2, 4, 3)$	$\{5\}$

(p_1, p_2, p_3)	$\{i : \pi_i \in \mathcal{ASC}_{(p_1,p_2,p_3)}\}$
$(3, 1, 2)$	$\{1\}$
$(3, 1, 4)$	$\{1\}$
$(3, 2, 1)$	$\{6\}$
$(3, 2, 4)$	$\{1\}$
$(3, 4, 1)$	$\{3\}$
$(3, 4, 2)$	$\{2\}$
$(4, 1, 2)$	$\{4\}$
$(4, 1, 3)$	$\{4\}$
$(4, 2, 1)$	$\{6\}$
$(4, 2, 3)$	$\{4\}$
$(4, 3, 1)$	$\{6\}$
$(4, 3, 2)$	$\{6\}$

Table 2.2: Explicit verification of the property of being completely 3-scrambling for $\mathscr{Q} = (\pi_1, \ldots, \pi_6)$ of Example 2.2.4.

Remark 2.2.5. *The property of being* completely *k-scrambling implies the property of being k-mixing. Moreover, notice that $N^{\mathrm{mix}}(n,k)+1 \leq N^*(n,k)$: In fact, when considering an optimal family of completely k-scrambling permutations and dropping an arbitrary permutation, it is still guaranteed that every herewith lost ascending sequence $(p_1,\ldots,p_k)$ is nevertheless represented at least by its alternative counterpart $(p_k,p_2,\ldots,p_{k-1},p_1)$ in another permutation of the family.*

The next two propositions resemble easily derivable, useful properties of scrambling families. They can be found in the works [Dus50, Spe71, CCHZ13, BEI+12, Na21].

Proposition 2.2.6. *Let $n \geq k \geq 2$. Then, following assertions hold ($\mathscr{P}$ is a (completely) k-scrambling d-family of permutations of $[n]$).*

(i) Every family $\mathscr{P}'$ being a reordering of the elements of $\mathscr{P}$ is (completely) k-scrambling.

(ii) For each $q \in S_n$, also the family $\mathscr{P}_q := \{\pi \circ q : \pi \in \mathscr{P}\}$ is (completely) k-scrambling. ($\mathscr{P}_q$ results from subjecting the domain of the permutations in $\mathscr{P}$ to a relabeling.)

(iii) Let $\rho \in S_n$ denote the reversing permutation, i.e., $\rho(i) := n-i+1$. Then, the family $\{\rho \circ \pi : \pi \in \mathscr{P}\}$ resulting from relabeling symbols via ρ, is (completely) k-scrambling, too.

(iv) If $\mathscr{F}$ is completely k-scrambling, it is as well k-scrambling. Therefore, the relation $N(n,k) \leq N^*(n,k)$ is generally valid.

(v) $N(n,2) = N^*(n,2) = 2$, as $\mathscr{P} = (\pi, \pi \circ \rho)$ is always completely 2-scrambling (ρ defined as in (iii)).

(vi) The bounds $k \leq N(n,k) \leq n$ and $k! \leq N^*(n,k) \leq \binom{n}{k}k!$ apply. □

The previous result permits to assume (without loss of generality) that the first permutation is always the identity of S_n.

Proposition 2.2.7 (Monotonicity, [Dus50])**.** *Let $\mathscr{P} \subseteq S_n$ be (completely) k-scrambling and let $1 \leq \ell \leq k$ and $k \leq m \leq n$. Then, the following assertions are true. They imply that $N(n,k)$ and $N^*(n,k)$ both increase monotonically in n as well as in k.*

(i) The family $\mathscr{P}\restriction_{[m]} := \{(\mathrm{rank}(\pi(i), \pi([m])))_{i=1}^m : \pi \in \mathscr{P}\} \subseteq S_m$ is (completely) k-scrambling.

(ii) $\mathscr{P}$ is in particular (completely) ℓ-scrambling. □

We now focus on the class of *Sequence Covering Arrays* (SCAs), which was introduced in [KHL+12] without pointing out a connection to scrambling permutations. The following definition is later illustrated (see Example 2.3.7).

Definition 2.2.8. *A d-family $\mathscr{A} \subseteq S_n$ is called* sequence covering array over the alphabet $[n]$ of strength k, *if for every $(s_1, \ldots, s_k) \in S_{n,k}$ there is one permutation $\pi \in \mathscr{A}$ such that*

$$\pi^{-1}(s_1) < \pi^{-1}(s_2) < \ldots < \pi^{-1}(s_k). \tag{2.3}$$

Abbreviating, we say that $\mathscr{A}$ appertains to the class $\mathrm{SCA}(d, n, k)$, *or less specifically, if the cardinality is not of interest, to the class* $\mathrm{SCA}(n, k)$. *Every permutation of $\mathscr{A}$, in which a fixed $s \in S_{n,k}$ fulfills* (2.3) *is said to* cover *s; less specifically we also say that $\mathscr{A}$ covers s, provided a covering permutation exists.*

Regarding a family of permutations as matrix, the latter definition means that every string of length k (containing no duplicates) is locatable in at least one row of the matrix as subsequence.

The following basic observation is crucial. It translates properties of monotonicity to properties of localizability of subsequence patterns.

Lemma 2.2.9 ([CCHZ13])**.** *Let $n \geq k \geq 2$ be fixed. Let $\iota : S_n \to S_n$, $\pi \mapsto \pi^{-1}$. For every completely k-scrambling d-family $\mathscr{F} \subseteq S_n$, the family $\{\iota(\pi) : \pi \in \mathscr{F}\}$ appertains to* $\mathrm{SCA}(d, n, k)$. *Conversely, the image of every $\mathscr{A}$ of type* $\mathrm{SCA}(d, n, k)$ *under the map ι constitutes a completely k-scrambling d-family.*

Proof. First, we show that monotonicity implies containment as subsequence: If $\pi(x_1) < \ldots < \pi(x_k)$, then $\iota(\pi) = (\pi^{-1}(1), \ldots, \pi^{-1}(n))$ in particular contains

$$(\pi^{-1}(\pi(x_1)), \pi^{-1}(\pi(x_2)), \ldots, \pi^{-1}(\pi(x_k)) = (x_1, \ldots, x_k)$$

as subsequence.

For the other proof direction, if π does not satisfy $\pi(x_1) < \pi(x_2) < \ldots < \pi(x_k)$, then there must be $\psi \in S_k \setminus \{\mathrm{id}\}$ such that $\pi(x_{\psi(1)}) < \ldots < \pi(x_{\psi(k)})$, which implies (as before) that $\iota(\pi)$ contains $(x_{\psi(1)}, \ldots, x_{\psi(k)})$ as subsequence. Consequently $(x_1, \ldots, x_k)$ is not a subsequence of $\iota(\pi)$. □

Remark 2.2.10. *When introducing completely scrambling permutations in [Spe71], Spencer presented them in a subtly different fashion: He considered them as families of linear orders on the set $[n]$ such that any selection $(p_1, \ldots, p_k) \in S_{n,k}$ forms a monotonically increasing sequence with respect to at least one order of the family. He examined (non-completely) scrambling families from that point of view, too.*

Due to the isomorphism in Lemma 2.2.9, the quantity $N^*(n, k)$ coincides with the so-called *sequence covering array number* (being the minimum possible cardinality for which the class $\mathrm{SCA}(d, n, k)$ is non-empty, cf. [BEI$^+$12]).

Remark 2.2.11 ([CCHZ13])**.** *Notice that simply taking the permutations of a completely k-scrambling family does not provide a sequence covering array. The instance $\mathscr{2}$ in Example 2.2.4, not containing the subsequence $(3, 4, 2)$, demonstrates this.*

2.3 Related combinatorial structures

2.3.1 Covering arrays

We fix the terminology for covering arrays (abbreviated as CAs) as they are related to SCAs in a "broader sense". It seems they were responsible for the terming of SCAs in [KHL+12] (we will return to this when discussing applications in Chapter 7). On the other hand, there are some interconnections of theoretical nature between CAs and SCAs (cf. Lemma 4.2.17).

Definition 2.3.1 (Covering Array, [Slo93])**.** *Let $d, n, k, q \in \mathbb{N}^{\times}$ with $k \leq n$ and consider the alphabet $B_q = \{0, 1, \ldots, q-1\}$. A* covering array with configuration (n, k, q) *is a $d \times n$ matrix A over the alphabet B_q, such that in each $d \times k$ submatrix,*[1] *any possible k-tuple in B_q^k is present as at least one row. The parameter k is hereby called* strength *and q the* level*. For $q = 2$, we speak about* binary *covering arrays.*

Example 2.3.2. *The subsequent determines a binary covering array with $n = 4$, $k = 3$ on $d = 8$ rows. We obtain this example by pasting as rows the the words in $\{0,1\}^3$ into the lefter-most 8×3 submatrix and afterwards populating the last column exhaustively until no violation of Definition 2.3.1 occurs. After deletion of any column, any binary word of length three is present among the rows of the resulted matrix.*

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Definition 2.3.3 (Covering Array Number)**.** *For given n, k, q, the minimum cardinality d for which a $d \times n$ covering array of strength k and level q exists is called* covering array number *and we write* $\mathrm{CAN}(n, k, q)$*. For $q = 2$ we denote the latter number by* $\mathrm{CAN}_{\mathrm{bin}}(n, k)$*.*

Remark 2.3.4. *Foundational work related to binary CAs can be found in [Mar48] in the context of measure theory (the notion of independence of sets is introduced) and in [Nec65] for problems related to gating circuits. The n columns of a $d \times n$ binary CA can be interpreted as n indicator functions $[d] \to \{0,1\}$, i.e., as n subsets of $[d]$. Letting $U_1, \ldots, U_n$ be these subsets, the analogue of strength k, called k-*independence *(following [KS73]), requires to satisfy the following condition: For any k-set $I \subseteq [n]$, for any of its*

[1] Submatrix is understood as a not necessarily contiguous submatrix determined by column indices $j_1 < \ldots < j_k$.

subsets $A \subseteq I$*,*

$$\bigcap_{i \in A} U_i \cap \bigcap_{j \in I \setminus A} ([d] \setminus U_j) \neq \emptyset.$$

2.3.2 Imposing regularity constraints

We discuss a regularized variant, first of completely scrambling families, then of scrambling families. Ignoring the fact that several structures were introduced again in different contexts, in different terminology, and in isomorphic forms, our following discourse shall reflect as much as possible the chronological order of when these families were first considered.

First we consider a particular class of block designs from *design theory*. According to [GG94], the first outlines[2] of this theory date back to the 19th century and are due to the Swiss geometer Jacob Steiner. In the course of the last century, this theory has experienced massive growth and was enriched by numerous facets. This is manifested in numerous publications (we refer here to the textbook [BJL99] and references therein), which combine various branches of mathematics, e.g. projective/affine geometry among many others. For the purposes of this thesis it seems most convenient to use the terminology of the recent paper [Yus20], with the subtle difference that, for consistency, we consider its objects as families of permutations (and not as multisets of permutations).

We now define perfect sequence covering arrays (PSCAs) being a particular block design of major interest for us.

Definition 2.3.5 (PSCA, [Yus20])**.** *Let* $\lambda \in \mathbb{N}^\times$*. By* $\mathrm{PSCA}(n, k, \lambda)$ *we refer to the class consisting of all families* $\mathscr{A} \in \mathrm{SCA}(n, k)$ *such that furthermore each subsequence* $s \in S_{n,k}$ *is covered by exactly* λ *permutations of* $\mathscr{A}$*. The parameter* λ *is hereby called* multiplicity*. Less specifically we refer to the class*

$$\mathrm{PSCA}(n, k) := \bigcup_{\lambda \in \mathbb{N}^\times} \mathrm{PSCA}(n, k, \lambda).$$

PSCAs are therefore nothing else than SCAs covering all $s \in S_{n,k}$ *equally* often – they are more *regular* than arbitrary SCAs. PSCAs do not exist for all choices of (n, k, λ) as will be clarified soon.

Remark 2.3.6. *An immediate property of PSCAs is that their cardinality must necessarily be equal to* $\lambda k!$*; this is why in the expression* $\mathrm{PSCA}(n, k, \lambda)$ *we do not provide an additional parameter specifying the cardinality. Moreover, the property of being a* $\mathrm{SCA}(n, k)$ *counting* $k!$ *permutations is equivalent to being a* $\mathrm{PSCA}(n, k, 1)$ *(cf. [CCHZ13]) – per permutation, namely,* $\binom{n}{k}$ *subsequences are covered (in total* $\binom{n}{k}k! = |S_{n,k}|$ *subsequences have to be covered once).*

[2] These outlines can be found in the work [Ste53] of 1853.

Example 2.3.7. *In the following, the left matrix determines a* PSCA$(4,3,1)$*, in particular an optimal* SCA$(4,3)$*. Swapping the third and fourth column makes the tuple* $(1,4,2) \in S_{4,3}$ *uncovered, consequently destroys the SCA-property – in particular perfectness. Generally, violation of perfectness can be testified more easily, as just one sequence covered more often than allowed (*λ *times) has to be indicated. Below, indication of e.g.* $s = (1,2,4)$*, covered twice, suffices to deny perfectness.*

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 1 & 4 & 2 \\ 3 & 2 & 4 & 1 \\ 4 & 1 & 3 & 2 \\ 4 & 2 & 3 & 1 \end{bmatrix} \xrightarrow{\textit{column swap } (3\leftrightarrow 4)} \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{4} & 3 \\ 2 & 1 & 3 & 4 \\ 3 & \mathbf{1} & \mathbf{2} & \mathbf{4} \\ 3 & 2 & 1 & 4 \\ 4 & 1 & 2 & 3 \\ 4 & 2 & 1 & 3 \end{bmatrix}$$

The following is the analogue to Proposition 2.2.6 (i)-(iii).

Lemma 2.3.8 (adapted from [Na21])**.** *The class* PSCA(n,k,λ) *is closed under the following operations:*

(i) Rearranging the elements of the family.

(ii) Relabeling of symbols: Every permutation π *of the family is replaced by* $q \circ \pi$*, where* $q \in S_n$ *is fixed. In matrix notation this means that the symbols* $x \in [n]$ *are subjected to a simultaneous (throughout the entire matrix) substitution* $x \mapsto q(x)$*.*

(iii) Reversing the order of positions: Every permutation π *of the family is replaced by* $\pi \circ \rho$ *(*ρ *is the reversing permutation defined in Proposition 2.2.6). In matrix notation this corresponds to reversing the order of columns.*

□

Definition 2.3.9. *[Na21] Two instances of* PSCA(n,k,λ) *are* equivalent *if the first can be transformed to the second by application of the operations (i)-(iii) from Lemma 2.3.8.*

The next observation follows readily by induction. It quantifies the increase of multiplicity for a PSCA when it is interpreted as a PSCA of lower strength.

Lemma 2.3.10 (Nesting property, [Yus20])**.** *Let* S *be a* PSCA(n,k,λ)*. Then,* S *is automatically a* PSCA$(n,\ell,\lambda\frac{k!}{\ell!})$ *for every strength* $\ell \in [k]$*.* □

Remark 2.3.11. *Historically, even before PSCAs have become trendy, see e.g. [Yus20, Na21, GW22]), the focus was on the study of* directed packings*,* directed coverings*, and* directed designs *(cf. [Lev91, MvT99, CCHZ13]). Except for notational differences, given* $n,w,k,\lambda \in \mathbb{N}^{\times}$*, these represent matrices whose rows consist of tuples of* $S_{n,w}$ *such that each* $s \in S_{n,k}$ *is covered by at least (covering), by at most (packing), or by exactly (design)* λ *rows. For* $\lambda = 1$*, a directed design is often termed* Steiner system*.*

Let us also discuss how regularity was indirectly imposed on the isomorphic structure of completely scrambling families of permutations. We provide the original definition,[3] and show that it does nothing else than imposing regularity on completely scrambling permutations as we have already seen it in the setting of SCAs. We rely on the notion of the rank defined in Section 2.1.

Definition 2.3.12 ([ITT00, TIT03])**.** *A non-empty family $\mathscr{F} \subseteq S_n$ is called* k-rankwise independent*, if for each k-set $X = \{x_1, \ldots, x_k\}$ and k distinct values $r_1, \ldots, r_k \in [k]$, we have*

$$\Pr\left[\bigwedge_{i=1}^{k} \operatorname{rank}(\pi(x_i), \{\pi(x_1), \ldots, \pi(x_k)\}) = r_i\right] = \frac{1}{k!}, \tag{2.4}$$

when π is drawn uniformly at random from $\mathscr{F}$.

The condition (2.4) can be paraphrased by the following:

$$\Pr\left[\pi(x_1) < \pi(x_2) < \ldots < \pi(x_k)\right] = \frac{1}{k!} \tag{2.5}$$

The following observation better illuminates rankwise independence (it is a regularized variant of Lemma 2.2.9).

Lemma 2.3.13 (cf. also [Iur22])**.** *A family $\mathscr{F} \subseteq S_n$ is k-rankwise independent iff $\iota(\mathscr{F})$ is a PSCA of strength k ($\iota : S_n \to S_n,\ \pi \mapsto \pi^{-1}$).*

Proof. Let $\mathscr{F}$ be a k-rankwise independent d-family. Then, for an arbitrary position selection $p \in S_{n,k}$, there must exist precisely $\lambda := d/k!$ permutations in $\mathscr{F}$ being ascending along p. Consequently, by Lemma 2.2.9, there will be exactly λ permutations of $\iota(\mathscr{F})$ covering the tuple p. As $p \in S_{n,k}$ was arbitrary, $\iota(\mathscr{F})$ lies in $\mathrm{PSCA}(n, k, \lambda)$. The converse proof direction is shown analogously. □

The following observation tells us that a nesting property (analogous to the one for PSCAs, cf. Lemma 2.3.10) applies for k-rankwise independent families. Its validity was directly noticed in [TIT03] without pointing to design theory.

Corollary 2.3.14. *A k-rankwise independent family $\mathscr{P} \subseteq S_n$ is automatically ℓ-rankwise independent for $1 \leq \ell \leq k$.*

Proof. Convert $\mathscr{P}$ to a $\mathrm{PSCA}(n, k)$ via Lemma 2.3.13, use the nesting property of PSCAs Lemma 2.3.10, and map the structure interpreted as a $\mathrm{PSCA}(n, \ell)$ back via Lemma 2.3.13(ii) to a ℓ-rankwise independent family. □

[3] It should not baffle, that the definition is in probabilistic language, being a special case of *restricted min-wise independence* introduced in [BCFM00] in the context of randomness. We return to restricted min-wise independence later.

Let us now look at *k-restricted min-wise independence*, being older[4] than k-rankwise independence, and also being more general.[5] It was introduced without linking to scrambling permutations by [BCFM00]. It will turn out that it is de facto another kind of regularization of scrambling permutations.

A direct regularization of *k-scrambling* permutations leads to the following concept (the subsequent curly brackets are set intentionally to distinguish from Definition 2.3.16). We emphasize that in many works (cf. e.g. [BCFM00]) an arbitrary probability distribution on the family of permutations is considered. Our focus will be on the special case of uniform distribution, which is also the case getting the most attention in many works (cf. e.g. [BCFM00, ITT00, ITT03, TIT03]).

Definition 2.3.15 ($\{k\}$-restricted min-wise independence, [MS03])**.** *A non-empty family $\mathscr{F} \subseteq S_n$ is called* $\{k\}$-restricted min-wise independent*, if for a position selection $p \in S_{n,k}$, whenever a permutation from the family is randomly drawn (with uniform probability), we have*

$$\Pr\left[\pi \in \bigcap_{j=2}^{k} \mathcal{ASC}_{(p_1,p_j)}\right] = \Pr\left[\pi(p_1) = \min\left\{\pi(p_j) : j = 1,\ldots,k\right\}\right] = \frac{1}{k}. \tag{2.6}$$

The probability of $1/k$ indicates that for given $p = (p_1,\ldots,p_k)$ the left hand side of (2.6) remains invariant under a single swap of p_1 and p_j, $j = 1,\ldots,k$ (there are k swaps). Here we obtained a balancing property, which is a pre-stage of the sought-after and even more regular property stated below in Definition 2.3.16. We now state the special case of higher regularity (it allows to select *up to* k elements, instead of precisely k elements).

Definition 2.3.16 (k-restricted min-wise independence,[BCFM00])**.** *A non-empty family $\mathscr{P} \subseteq S_n$ is* k-restricted min-wise independent*, if for an arbitrary set $X \subseteq [n]$ with $|X| \leq k$ and an arbitrarily distinguished $x \in X$, the following condition is fulfilled: Whenever π is chosen uniformly at random from $\mathscr{P}$, we have*

$$\Pr\left[\min(\pi(X)) = \pi(x)\right] = \frac{1}{|X|}. \tag{2.7}$$

In case $n = k$, we abbreviate n-restricted min-wise independence just by min-wise independence *(as there is in fact no more restriction on X).*

It is immediate that the latter property satisfies a nesting property analogous to the one for k-rankwise independent permutations: If $\mathscr{F}$ is k-restricted min-wise independent, then it is as well ℓ-restricted min-wise independent, for $\ell \in [k]$.

The following observation tells us that the notion of restricted min-wise independence and rankwise independence (see Definitions 2.3.12 and 2.3.16) collapses for $k = 3$. The assertion was mentioned in [ITT00]; we add a short proof.

[4]Introduced in 1998 in [BCFM00]. However, according to [Ind01] it independently appears earlier in 1994 in the book [Mul94] for the purpose of randomness in computational geometry.

[5]Introduced in 2000 in [ITT00].

Lemma 2.3.17. *Let $\mathscr{P} \subseteq S_n$ be 3-restricted min-wise independent. Then, $\mathscr{P}$ is also 3-rankwise independent (and therefore isomorphic to a PSCA of strength 3).*

Proof. Set $d = |\mathscr{P}|$. Assume 3-restricted min-wise independence for $\mathscr{P}$ and that there exists $(p_1, p_2, p_3) \in S_{n,3}$ such that $\mathcal{ASC}_{(p_1,p_2,p_3)}$ is underrepresented among the members of $\mathscr{P}$ in the sense of (2.5), i.e., $|\mathscr{P} \cap \mathcal{ASC}_{(p_1,p_2,p_3)}| < d/3!$. Let us replace the permutations of $\mathscr{P}$ by their restrictions to $\{p_1, p_2, p_3\}$, and convert these restricted functions in an order preserving manner to permutations of $\{1, 2, 3\}$ (cf. Proposition 2.2.7). The end product will necessarily be a 3-restricted min-wise independent family $\mathscr{Q}$ of permutations in S_3 with under-representation of, say, $\mathcal{ASC}_{(1,2,3)}$ (without loss of generality). Moreover, by min-wise independence, up to reordering of elements,

$$\mathscr{P} = (\alpha_1, \ldots, \alpha_{d/3}, \beta_1, \ldots, \beta_{d/3}, \gamma_1, \ldots, \gamma_{d/3}), \tag{2.8}$$

where for $j = 1, \ldots, d/3$, we have $\alpha_j(1) = 1$, $\beta_j(2) = 1$, and $\gamma_j(3) = 1$. The underrepresentation of $\mathcal{ASC}_{(1,2,3)}$ implies an underrepresentation of $\mathcal{ASC}_{(2,3)}$, meaning $|\mathscr{P} \cap \mathcal{ASC}_{(2,3)}| < d/2$. This follows immediately from the fact that $|(\beta_j)_j \cap \mathcal{ASC}_{(2,3)}| + |(\gamma_j)_j \cap \mathcal{ASC}_{(2,3)}| = d/3+0$. We obtain a contradiction, as for being min-wise independent, it is a necessity to have a balanced representation of $\mathcal{ASC}_{(2,3)}$ and $\mathcal{ASC}_{(3,2)}$. □

Min-wise independence has been studied also in a form tolerating a relative error.

Definition 2.3.18 ([BCFM00])**.** *Let $\varepsilon \geq 0$. A non-empty family $\mathscr{F} \subseteq S_n$ is called* ε-approximately k-restricted min-wise independent*, if for any $X \subseteq [n]$ with $|X| \leq k$, any choice of $x \in X$ implies*

$$\left| \Pr\left[\pi(x) = \min \pi(X)\right] - \frac{1}{|X|} \right| \leq \frac{\varepsilon}{|X|},$$

for a random choice of $\pi \in \mathscr{F}$ that follows the uniform probability distribution on $\mathscr{F}$. If $k = n$, we omit the infix "k-restricted".

Let us conclude the chapter with the categorization and an overview of the fallen concepts. The notation and approach of the chapter was chosen to accomplish this (e.g. we do not consider sets of permutations as some authors do). From Figure 2.1 we can read off which properties on permutation families are implied by other properties. At first, one might be tempted to suspect that some properties are so strong that, especially when combined with others, they could potentiate to even stronger properties. It turns out, as we will clarify, that these first intuitions cannot be confirmed in general (observe the non-vanishing intersections and differences in the Venn diagram of Figure 2.1 and consider the counterexamples in Remark 2.3.19).

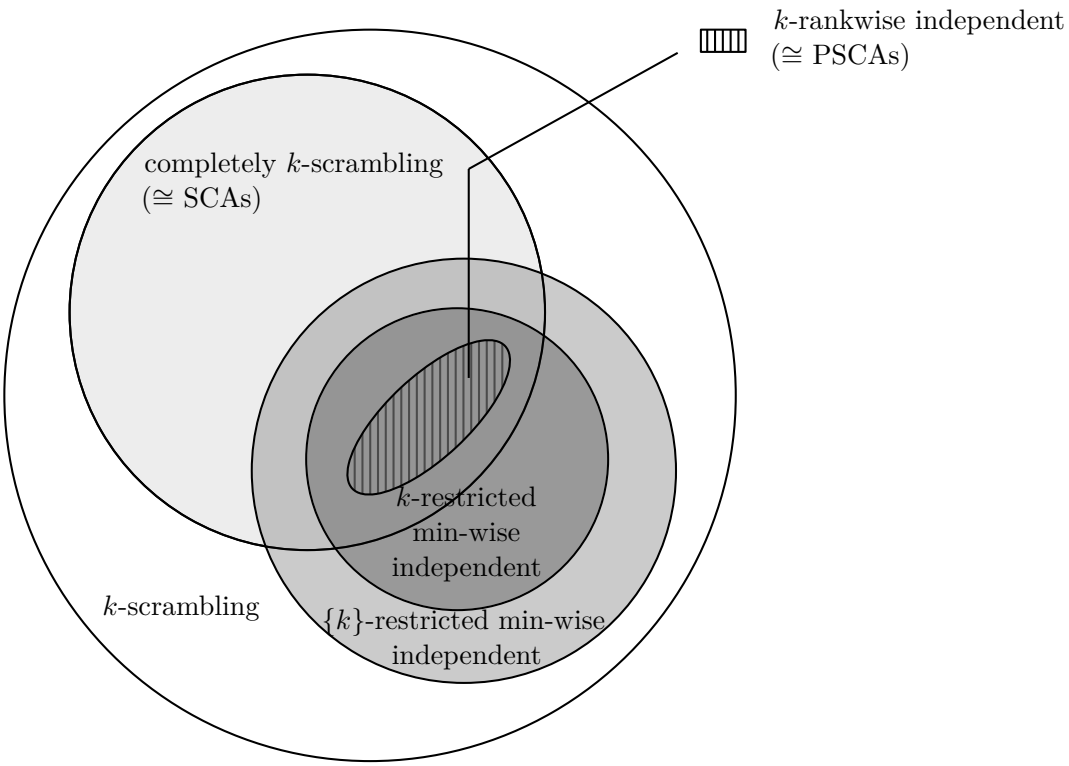

Figure 2.1: Venn diagram classifying the various properties of permutation families (with fixed n and k, both of general character).

Remark 2.3.19. *The permutation family* $\mathscr{O} = \{(1,2,3,4),(3,2,1,4),(3,2,4,1)\}$ *serves as an example of a* $\{3\}$*-restricted min-wise independent while simultaneously not* $\{2\}$*-restricted min-wise independent family.*[6] *Hence, in contrast to* k*-rankwise independent families, for which the nesting (for decreasing* k*) is automatically implied, here, nesting can only be ensured by explicitly enforcing it. This enforcement leads to the notion of* k*-restricted min-wise independence.*

Restricted min-wise independent families additionally being completely scrambling are not necessarily rankwise independent, as can be seen by inspecting a family of $d = 36$ *permutations (with strength* $k = n = 4$*) comprising the entire group* S_4 *and additionally the elements from* $\{\pi \in S_4 : \pi^{-1}(2) < \pi^{-1}(3)\}$ *(incorporated as duplicates).*

[6] We can namely observe $|\mathscr{O} \cap \mathcal{ASC}_{(1,2)}| = 1$ and $|\mathscr{O} \cap \mathcal{ASC}_{(2,1)}| = 2$. Even faster, this can be seen by $\operatorname{lcm}(\{1,2,3\}) \nmid |\mathscr{O}|$.

CHAPTER 3

Methods

3.1 The probabilistic method

We emphasize in advance that several existence proofs appearing in this thesis are based on a method which allows to formulate proofs for the existence of a combinatorial structure in a very concise and elegant way. The basic idea is that we endow a finite set of combinatorial structures with a suitable probability distribution (usually with the uniform distribution) and afterwards we try to show that a random variable sampled according to the distribution corresponds with positive probability to one of those structures with the desired properties. If we can show positivity, we automatically can conclude that at least one such structure must exist. We observe that this method can yield existence of a desired structure but is not capable of providing a witnessing instance. This philosophy of proving is often called the *probabilistic method* [AS00]. According to [AS00] this argumentation technique should be attributed to Erdős, as frequently employed by him.

The method seems to be useful especially for the so-called branch of *extremal combinatorics*. To clarify what is here meant by *extremal*, we refer to the description given in [Alo03]:

> "Extremal combinatorics deals with the problem of determining or estimating the maximum or minimum possible cardinality of a collection of finite objects that satisfies certain requirements. Such problems are often related to other areas including computer science, information theory, number theory and geometry. This branch of combinatorics has developed spectacularly over the last few decades [...]."

For an arbitrary probability space U, let $A = \{a_1, \ldots, a_N\} \subseteq U$ collect events with associated probabilities $p_1, \ldots, p_N$. For a function $f : A \to \mathbb{R}$ we sometimes use, for

referring to the conditional expectation $\mathbb{E}[f(x)|x \in A]$, the notation

$$\mathop{\mathbb{E}}_{a\in A}[f(a)] = \frac{1}{\sum_{a\in A}\Pr[a]}\sum_{a\in A}\Pr[a]\,f(a). \tag{3.1}$$

For some proofs relying on the probabilistic method, it will be sufficient to employ the brute *union bound of probability*, also known as *Boole's inequality/subadditivity for (probability) measures*, which affirms that ($I \subseteq \mathbb{N}$)

$$\Pr\left[\bigcup_{i\in I} A_i\right] \leq \sum_{i\in I}\Pr[A_i].$$

We will also fall back on the following bound that quantifies a relative error.

Theorem 3.1.1 (Binomial version of Chernoff bound, [AS00, p. 268])**.** *Let $X_1, \ldots, X_N$ independent random variables attaining the value 1 with probability p_i and attaining the value 0 with probability $1 - p_i$, $i = 1, \ldots, N$. Let $Y = \sum_{i=1}^{N} X_i$ and $\mu = \mathbb{E}[Y]$. For any $\varepsilon > 0$, there is a constant $c_\varepsilon > 0$, depending only on ε, such that*

$$\Pr[|Y - \mu| > \varepsilon\mu] < 2e^{-c_\varepsilon \mu}. \tag{3.2}$$

□

The value in (3.2) for c_ε can be chosen (cf. [AS00]) as

$$c_\varepsilon = \min\left(-\ln(e^\varepsilon(1+\varepsilon)^{-(1+\varepsilon)}), \varepsilon^2/2\right) \geq \varepsilon^2/3,$$

which leads to the explicit, weakened form of (3.2),

$$\Pr[|Y - \mu| > \varepsilon\mu] \leq 2e^{-\varepsilon^2\mu/3}. \tag{3.3}$$

3.2 Information theory for extremal combinatorics

In this section we resemble some important properties of the (Shannon) entropy. We follow the presentation of the textbook [CT06]. Entropy methods have proven to be a powerful tool also in combinatorics, however, more classical application domains are coding/communication/quantum information theory, data compression, cryptography, statistics, thermodynamics, and computational molecular biology (cf. [CT06] and references therein). Katona's paper [Kat66] is an early work applying entropy methods in the context of combinatorics (it seems, by the way, to be the first work of this kind [Für96]).[1] In [Alo03] a selection of problems from extremal combinatorics solved with such entropy methods is given.

[1] Katona used the entropy to provide a lower bound for so-called separating systems of sets [Kat66].

3.2.1 Information theory

We now discuss main concepts of the Shannon entropy, following the presentation of [CT06] – for a more extensive treatment of the topic, we point to the latter work.

In the subsequent we consider exclusively *discrete* random variables, i.e., possessing *finite range*. We denote the probability mass function of a random variable X by

$$p_X : \mathrm{range}(X) \to [0,1], \quad x \mapsto \Pr[X = x].$$

Definition 3.2.1. *The (binary) entropy of a random variable X is defined as*

$$H[X] := -\sum_{\substack{x \in \mathrm{range}(X) \\ p_X(x) > 0}} p_X(x) \log_2 (p_X(x)). \tag{3.4}$$

Values of the range occurring with zero probability do not have an impact on the entropy (in principle the range can be extended by arbitrarily many new elements occurring with zero probability without affecting the entropy). For brevity we will denote the sum in (3.4) without the restriction $p_X(x) > 0$ using just the convention $0 \cdot \log_2 (0) = 0$. The convention also permits to interpret the entropy of X as the expectation of the random variable $-\log_2 (p_X)$.

Intuitively, we think of the entropy as the expected number of bits needed to describe an object randomly drawn from a collection.

Example 3.2.2. *If a random variable attains precisely two values, say* $\mathrm{range}(X) = \{0,1\}$, *by setting* $p = p_X(1)$, *the identity* (3.4) *simplifies to*

$$H[X] = \mathsf{H}(p) := -p \log_2 (p) - (1-p) \log_2 (1-p). \tag{3.5}$$

The function $\mathsf{H}(p)$, *defined on the interval* $[0,1]$ *(with the aforementioned continuity convention at the boundary point* 0*), will appear in multiple contexts. It is concave and its graph possesses the symmetry axis* $\left\{(\frac{1}{2}, y) : y \in \mathbb{R}\right\}$. *The entropy of X is hence maximized for* $p = \frac{1}{2}$, *i.e., if p_X corresponds to the uniform distribution.*

The following assertion is also a consequence of the concavity of the logarithm.

Lemma 3.2.3. *Let X be a random variable with* $|\mathrm{range}(X)| = n$. *Then,* $H[X] \leq \log_2 (n)$ *and equality holds iff p_X is the uniform distribution.* □

Entropy can also be considered for a vector of discrete random variables (technically being nothing else than a single random variable possessing as attainable values all elements of the Cartesian product of ranges of the variables comprised by the vector).

Definition 3.2.4 (Joint entropy)**.** *Let $X_1, \dots X_k$ be discrete random variables. Their* joint entropy *is defined as*

$$H[X_1, \dots, X_k] = \sum_{(x_1,\dots,x_k)\in \text{range}(X_1)\times\cdots\times\text{range}(X_k)} p(x_1, \dots, x_k) \log_2 \left(p(x_1, \dots, x_k)\right), \quad (3.6)$$

with $p(x_1, \dots, x_k) = \Pr[X_1 = x_1 \wedge \dots \wedge X_k = x_k]$ denoting the joint probability mass function.

Definition 3.2.5 (Conditional entropy)**.** *Let X, Y be discrete random variables and assume that (X, Y) has joint probability mass function $p(x, y)$. The quantity*

$$H[Y|X] := \sum_{x\in\text{range}(X)} p(x) H[Y|X = x] = -\sum_{x\in\text{range}(X)} p(x) \sum_{y\in\text{range}(Y)} p(y|x) \log_2 \left(p(y|x)\right) \quad (3.7)$$

is called the entropy of Y conditioned to X*.*

Conditional and joint entropy obey the following law.

Theorem 3.2.6 (Chain rule)**.** *For discrete random variables X, Y, Z the following rules apply.*

(i) $H[X, Y] = H[X] + H[Y|X]$.

(ii) $H[X, Y|Z] = H[X|Z] + H[Y|X, Z]$.

□

Combining both asserts in Theorem 3.2.6 permits to derive a k-ary chain rule for the joint entropy of random variables.

Corollary 3.2.7. *For discrete random variables $X_1, \dots, X_k$ we have*

$$H[X_1, \dots, X_k] = \sum_{i=1}^{k} H[X_i|X_{i-1}, X_{i-2}, \dots, X_1] \quad (3.8)$$

□

Corollary 3.2.8 (Independence bound on entropy)**.** *Let $X_1, \dots X_k$ be discrete random variables with joint probability mass function $p(x_1, \dots, x_n)$. Then,*

$$H[X_1, \dots, X_k] \leq \sum_{i=1}^{k} H[X_i]. \quad (3.9)$$

□

3.3 Deletion correcting codes

In this section we resemble the terminology for deletion correcting codes and focus on useful aspects of the Varshamov-Tenengolts codes (introduced in [VT65], later analyzed among others in [Lev91]). These codes form an interesting instance of single deletion correcting codes. In the subsequent, we follow the terminology of [Lev91].

Definition 3.3.1. *For $q \in \mathbb{N}^\times$, consider the alphabet $B_q := \{0, \dots, q-1\}$. A t-tuple $w = (w_1, \dots, w_t) \in B_q^t$ is called a word (of length t). The Kleene closure of B_q is denoted by $B_q^* = \bigcup_{t \geq 0} B_q^t$.*

Definition 3.3.2. *For a word $w \in B_q^t$ let $\|w\| := \sum_{j=1}^t w_j$ (corresponding for $q = 2$ to the Hamming distance between w and the zero tuple of B_q^t.). Moreover, let us define its position-weighted variant $W(w) := \sum_{j=1}^t j w_j$.*

Definition 3.3.3. *Under a code C we understand a subset of B_q^*. Its elements are called codewords. For $s \in \mathbb{N}^\times$ and a length-t word w, denote the set of all words resulting from w after s deletions by*

$$\mathsf{del}_s(w) := \left\{ x \in B_q^{t-s} : x \text{ is a subsequence of } w \right\}.$$

For a set of words $\mathfrak{W}$, define $\mathsf{del}_s(\mathfrak{W}) := \{\mathsf{del}_s(w) : w \in \mathfrak{W}\}$. A code $C \subseteq \mathfrak{W} \subseteq B_q^*$ is said to be an s-covering from below *of $\mathfrak{W}$ if $\mathsf{del}_s(C) = \mathsf{del}_s(\mathfrak{W})$.*

The following definition could be posed more generally to deal as well with *insertion correcting codes* (cf. [Lev91]).

Definition 3.3.4 ([Lev91])**.** *Let $C \subseteq \mathfrak{W} \subseteq B_q^n$ (containing only equally long words). C is called a* $\mathfrak{W}$-perfect code capable of correcting s deletions *if C is a s-covering from below of $\mathfrak{W}$.*

Definition 3.3.5 (Varshamov–Tenengolts codes, [VT65])**.** *For $\ell, m \in \mathbb{N}^\times$ there is a unique decomposition $\ell = ms + r$ with $s \in \mathbb{Z}$ and $r \in \{0, \dots, m-1\}$; denote that residual r by $r_m(\ell)$. Define the* Varshamov–Tenengolts codes *as the $n+1$ codes given by*

$$\mathrm{VT}^{n,a} := \{w \in B_2^n : r_{n+1}(W(w)) = a\}, \quad a = 0, \dots, n. \tag{3.10}$$

Example 3.3.6. *The space of words B_2^4 contains the* 16 *words (denoted as strings)* `0000`,`0001`,...,`1110`,`1111`. *In Figure 3.1 the* 5 *Varshamov-Tenengolts codes partitioning the set of these words are displayed.*

Theorem 3.3.7 ([Lev91])**.** *Consider B_2^n for $n \in \mathbb{N}^\times$. Then, all Varshamov-Tenengolts codes $\mathrm{VT}^{n,a}$, $a = 0, \dots, n$, are B_2^n-perfect codes capable of correcting one deletion.*

Proof. Fix n and $a \in \{0, \dots, n\}$. The following will to be shown: For each $u = (u_1, \dots, u_{n-1}) \in B_2^{n-1}$, there is a word $w(n, a, u) \in \mathrm{VT}^{n,a} \subseteq B_2^n$ satisfying $u \in \mathsf{del}_1(w)$.

$a=0$	$a=1$	$a=2$	$a=3$	$a=4$
0000	0101	0011	0010	0001
0110	1000	0100	1011	0111
1001	1110	1101	1100	1010
1111				

Figure 3.1: The codes $\mathrm{VT}^{4,a}$, $a = 0, \ldots, 4$ partitioning $\{0,1\}^4$.

The latter appertainance holds iff it is always possible to determine an appropriate index position $i \in [n]$ and a value $v \in B_2$ such that

$$w = (u_1, \ldots, u_{i-1}, v, u_i, \ldots, u_{n-1}) \in \mathrm{VT}^{n,a}\,. \tag{3.11}$$

The latter requirement is satisfied iff there exists an appropriate scalar $k \in \mathbb{Z}$ such that

$$k(n+1) + a = W(w) = \sum_{j=1}^{i-1} j u_j + iv + \sum_{j=i}^{n-1} (j+1) u_j \tag{3.12}$$

$$= W(u) + iv + \sum_{j=i}^{n-1} u_j. \tag{3.13}$$

Rearranging, we obtain

$$k(n+1) + a - W(u) = \sum_{j=i}^{n-1} u_j + iv \in \begin{cases} \{\|u\|+1, \ldots, n\}\,, & \text{iff } v = 1 \\ \{0, \ldots, \|u\|\}\,, & \text{iff } v = 0 \end{cases}. \tag{3.14}$$

where the equality *can* only hold when

$$v = 1 \iff r := r_{n+1}(a - W(u)) \in \{\|u\|+1, \ldots, n\}\,. \tag{3.15}$$

Consequently, given the information u, regardless of the position of insertion i, it is clear how the letter v must necessarily be defined to satisfy (3.11). We now argue why there always exists a suitable position choice for i, meaning that the following equality is satisfiable

$$\sum_{j=i}^{n-1} u_j + iv = r_{n+1}(a - W(u)) \in \{0, \ldots, n\}\,. \tag{3.16}$$

From (3.16) it becomes clear what necessary choices for i are: For definiteness, let us take the largest index among the feasible indices for i and define

$$i(n,a,u) := \begin{cases} \max\left\{ i \in [n] : \sum_{j=i}^{n-1} u_j = r_{n+1}(a - W(u)) \right\}, & r \in \{0, \ldots, \|u\|\} \\ \max\left\{ i \in [n] : \sum_{j=i}^{n-1} u_j = r_{n+1}(a - W(u) - i) \right\}, & r \in \{\|u\|+1, \ldots, n\} \end{cases}. \tag{3.17}$$

The maxima are well-defined, as taken over non-empty sets: Indeed, in case of $v = 0$, as the involved sum attains for $i = 1$, $i = n$ the values $\|u\|$, 0 respectively, there is at least one choice for i such that the sum attains any value $r \in \{0, \ldots, \|u\|\}$. In case of $v = 1$, the sum attains the same range of values as before, while contemporarily $r_{n+1}(a - W(u) - i)$, $i = 1, \ldots, n$, traverses in reversed order the same range of numbers; this implies existence of a collision value i, for which equality holds. The proof is completed, as for $u \in B_2^{n-1}$, the composed word

$$w := (u_1, \ldots, u_{i(n,a,u)-1}, v(n, a, u), u_{i(n,a,u)}, \ldots, u_n), \tag{3.18}$$

where $v(n, a, u) \in \{0, 1\}$ with $v(n, a, u) = 1$ iff $r_{n+1}(a - W(u)) \in \{\|u\| + 1, \ldots, n\}$, is a member of $\mathrm{VT}^{n,a}$. □

Corollary 3.3.8 ([Lev91])**.** *Given fixed numbers $0 \leq a \leq n$, each word in $w \in B_2^{n-1}$ has a* unique *"correcting" codeword $c \in \mathrm{VT}^{n,a}$, i.e., $w \in \mathsf{del}_1(c)$. Moreover, the codes $\mathrm{VT}^{n,a}$ form a partition of B_2^n.*

Proof. The partitioning property is clear (subdivision in congruence classes). For uniqueness, recalling the previous proof of Theorem 3.3.7, the only thing left to prove is that the feasible values for i, i.e., those over which the maximum is taken in (3.17) lead to the same composition in (3.18).

Depending on its inserted value in (3.18), we make a case distinction and first consider $v(n, a, u) = 0$: Let $i < l$ such that

$$\sum_{j=i}^{n-1} u_j = \sum_{j=l}^{n-1} u_j = r_{n+1}(a - W(u)).$$

Consequently, $u_i, \ldots, u_{l-1}$ must be zeros and it is indifferent for the composition before which of the positions $i, \ldots, l$ the novel zero is placed.

The remaining case $v(n, a, u) = 1$ meets an analogous indifference. □

Remark 3.3.9. *Let us briefly explain the naming, i.e., point out in which sense these $\mathrm{VT}^{n,a}$-codes have the ability to* correct deletions*: Suppose a message M residing in a set of messages $\{M_1, \ldots, M_m\}$ has to be transmitted from device A to device B (the set is known for A and B). Furthermore, assume that the communication channel between A and B is only prone to the following type of disturbance: The potential loss of one bit of information during transmission of ℓ bits, where the position of the lost bit is unknown.*

In advance, we can choose n large enough such that a suitable $a \in \{0, \ldots, n\}$ exists, for which $\{M_1, \ldots, M_m\}$ is injectively embeddedable in $\mathrm{VT}^{n,a}$ (as before, n and a are known values for A and B).

Device A sends as an encoding of M its identifying codeword $w \in \mathrm{VT}^{n,a}$. After communication, if B has received a n-bits word, B can just map it back to the message space. Otherwise, i.e., if $n - 1$ bits have arrived, before this conversion, the uniquely

reconstructable n-bits word in $\mathrm{VT}^{n,a}$ has to be built by B in an intermediate, correcting step.

The cardinalities $|\mathrm{VT}^{n,a}|$, $a = 0, \ldots, n$, might vary (cf. Figure 3.1). For given n, this raises the question about the pattern of the sequences $(|\mathrm{VT}^{n,a}|)_{a=0}^{n}$, which are cataloged as entry A053633 in [SI22b].

Indeed, there is a remarkable representation as sum for $|\mathrm{VT}^{n,a}|$ due to [Gin67] implying furthermore that among the values $|\mathrm{VT}^{n,a}|$, $a = 0, \ldots, n$, the largest cardinality is obtained for $a = 0$, whereas the smallest for $a = 1$ (cf. also Figure 3.1).

In [Slo08] a similar representation formula is derived. With φ denoting Euler's totient function and μ the Möbius function, it reads as follows.

Theorem 3.3.10 ([Slo08])**.** *Let $n \in \mathbb{N}^{\times}$. For each $a = 0, \ldots, n$, the cardinality of $\mathrm{VT}^{n,a}$ can be determined by the formula*

$$|\mathrm{VT}^{n,a}| = \frac{1}{2(n+1)} \sum_{2 \nmid d \,|\, (n+1)} \varphi(d) \frac{\mu\left(\frac{d}{\gcd(d,a)}\right)}{\varphi\left(\frac{d}{\gcd(d,a)}\right)} 2^{(n+1)/d}.$$

□

CHAPTER 4

Asymptotics

We discuss asymptotic bounds for (completely) scrambling permutations as well as for their regularizations. Not all bounds provide a way to actually construct such permutation families – this task is deferred to the next chapter.

4.1 k-scrambling permutations

In [Spe71] the following bounds are obtained.

Theorem 4.1.1 (Hajnal-Spencer)**.** *Let $n \geq k \geq 3$. We have (k fixed, $n \to \infty$)*

$$\log_2 \log_2 n \leq N(n,3) \tag{4.1}$$

$$\leq N(n,k) \leq k2^k(1+o_n(1)) \log_2 \log_2 n. \tag{4.2}$$

Before proving the bounds (4.1) and (4.2), we provide some preparatory observations.

Lemma 4.1.2 (Erdős-Szekeres, variant)**.** *For $m \in \mathbb{N}^\times$ and two permutations $\pi, \sigma \in S_{m^2+1}$, there exists a selection of $m+1$ positions $(p_1, \ldots, p_{m+1}) \in S_{m^2+1,m+1}$ such that $\pi(p_1), \ldots, \pi(p_{m+1})$ and $\sigma(p_1), \ldots, \sigma(p_{m+1})$ form monotone sequences.*

Proof. The classical Erdős-Szekeres theorem (see [ES35, p. 467]) affirms that any sequence of length m^2+1 possesses a monotone subsequence of length $m+1$. We can therefore identify $x_1 < \ldots < x_{m+1}$ such that $\pi(\sigma^{-1}(x_1)) \prec \pi(\sigma^{-1}(x_2)) \prec \ldots \prec \pi(\sigma^{-1}(x_{m+1}))$ with $\prec \in \{<, >\}$. Set $p_1 := \sigma^{-1}(x_1), \ldots, \sigma^{-1}(x_{m+1})$. As a consequence, $\pi(p_1) \prec \pi(p_2) \prec \ldots \prec \pi(p_{m+1})$ and $\sigma(p_1) < \sigma(p_2) < \ldots < \sigma(p_{m+1})$, i.e., π and σ are monotone along $p_1, \ldots, p_{m+1}$. □

Corollary 4.1.3. *For a family $\mathscr{F}$ consisting of $s+1$ permutations of $[2^{2^s}+1] \subseteq \mathbb{N}^\times$, there exist $(p_1, p_2, p_3) \in S_{2^{2^s}+1,3}$ such that $\pi(p_1), \pi(p_2), \pi(p_3)$ is a monotone sequence for every $\pi \in \mathscr{F}$.*

Proof sketch. We note that $2^{2^s}+1 = \left(2^{2^{s-1}}\right)^2 + 1$. Therefore with $m := 2^{2^{s-1}}$ we face a family of $s+1$ permutations each having length m^2+1. Consequently, by Lemma 4.1.2 the first two permutations will share a subsequence of positions of length $m+1 = \left(2^{2^{s-2}}\right)^2+1$ along which both permutations are monotonic. All the family's permutations can now be restricted to the indices of the subsequence. After the shrinkage one has to identify all these $s+1$ shrunken tuples in an order preserving manner with tuples in $S_{m+1,m+1}$, such that we deal with a $(s+1)$-family of permutations of $[m+1] = \left[\left(2^{2^{s-2}}\right)^2+1\right]$.

Then, the outlined process can be repeated in order to obtain a selection of indices along which the first three permutations of the family are monotone. This can be iterated further such that after in total s iterations one ends up with a triple of position indices along which all the $s+1$ members of the family constitute monotone sequences. □

Proof of the lower bound (4.1) *in Theorem 4.1.1.* The following simple scenario immediately negates the property of being 3-scrambling: If one can individuate a position selection $(p_1, p_2, p_3) \in S_{n,3}$ such that, for all members π of a permutation family, monotonicity is fulfilled along (p_1, p_2, p_3), i.e., either $\pi(p_1) < \pi(p_2) < \pi(p_3)$ or $\pi(p_1) > \pi(p_2) > \pi(p_3)$, then the family cannot be 3-scrambling, as it is impossible to find a permutation π satisfying $\pi(p_2) < \min(\pi(p_1), \pi(p_3))$.

Assume by contradiction that there exists a 3-scrambling family of permutations of $[n]$ consisting of $c := 2 + \lfloor \log_2 \log_2 (n-1) \rfloor$ members. Let now $s \in \mathbb{N}^\times$ be chosen such that

$$2^{2^{s-1}} + 1 \leq n < 2^{2^s} + 1. \tag{4.3}$$

By monotonicity (see Proposition 2.2.7), there must hence exist a 3-scrambling c-family $\mathscr{F}$ consisting of permutations of $[2^{2^{s-1}}+1]$. By the choice of s, the bound $c \leq s+1$ holds. By Corollary 4.1.3 we can find three index positions along which all permutations of $\mathscr{F}$ are monotone. This contradicts the property of being 3-scrambling. Lastly, c majorizes $\log_2 \log_2 n$ for $n \geq 3$. □

Turning our attention now on the upper bound, we need an auxiliary result concerning the boundedness of $\mathrm{CAN}_{\mathrm{bin}}(\cdot)$ and being of independent interest. We emphasize that there have been many successful efforts to quantify predominantly the upper bound of $\mathrm{CAN}_{\mathrm{bin}}$. We refer to [Slo93, LKL+11] for extended discussions and comparisons of results on related upper bounds in literature (cf. [Rou87] for a construction yielding the best bound for $k = 3$). We now provide a known bound being valid for arbitrary k. Hereby, we will notice a structural similarity ($k!$ gets replaced with 2^k) to Spencer's upper bound for *completely* k-scrambling permutations (see Proposition 4.2.1).

Theorem 4.1.4 ([KS73])**.** *Let $n \geq k$. Then as $n \to \infty$, for fixed k,*

$$\mathrm{CAN}_{\mathrm{bin}}(n,k) \leq \frac{k}{\log_2 \frac{2^k}{2^k-1}} (1 + o_n(1)) \log_2 n. \tag{4.4}$$

Proof. We make use of the probabilistic method. Populate a $d \times n$ matrix with values drawn from $\{0,1\}$ independently and uniformly at random. For the sampled matrix, the probability to fail to be a binary covering array is given by the following consideration: The probability that at least one requirement is violated, is majorized by summed probabilities of violations encountered at any constellation of an arbitrary binary word $w \in \{0,1\}^k$ coupled with an arbitrary position selection $p \in \binom{[n]}{k}$. Hence, there are $2^k\binom{n}{k}$ constellations where each respective probability of violation is $\left(1-\frac{1}{2^k}\right)^d$. Therefore, individuating a large enough value for d such that

$$T(d,n,k) := 2^k \binom{n}{k} \left(1-\frac{1}{2^k}\right)^d < 1, \tag{4.5}$$

proves the bound (4.4). □

Remark 4.1.5. *For the dual problem, asking for $d \in \mathbb{N}^\times$ to determine*[1]

$$\mathrm{CAK}_{\mathrm{bin}}(d,k) := \max\left\{n \in \mathbb{N}^\times : \ A\ d \times n \text{ binary CA of strength } k \text{ exists}\right\}, \tag{4.6}$$

we obtain that each small enough n satisfying (4.5) *certainly implies $n \leq \mathrm{CAK}_{\mathrm{bin}}(d,k)$. In particular, the values for n even granting $T^{\mathrm{approx}} := (2en/k)^k e^{-d2^{-k}} < 1$, imply $n \leq \mathrm{CAK}_{\mathrm{bin}}(d,k)$ (as by estimation*[2] *$T < T^{\mathrm{approx}}$). The latter sharper restriction on n means that $n < \frac{k}{2e} e^{d2^{-k}k^{-1}}$, consequently*

$$\mathrm{CAK}_{\mathrm{bin}}(d,k) \geq \left\lfloor \frac{k}{2e} e^{d2^{-k}k^{-1}} \right\rfloor > \frac{k}{2e} e^{d2^{-k}k^{-1}} - 1. \tag{4.7}$$

We are now ready for deriving the upper bound. It relies on a creative idea of Hajnal discussed in [Spe71, SW20], whose argumentation we take up. To be in line with other proofs appearing in this thesis, in the subsequent proof, however, a link to binary CAs of strength $(k-1)$ is preferred over a link to $(k-1)$-independent sets (cf. Remark 2.3.4).

Proof of the upper bound (4.2) *in Theorem 4.1.1.* Let

$$s := \min\left\{\tilde{s} \in \mathbb{N}^\times : 2^{\mathrm{CAK}_{\mathrm{bin}}(\tilde{s},k-1)} \geq n\right\} \tag{4.8}$$

and save the quantity $M := \mathrm{CAK}_{\mathrm{bin}}(s,k-1)$. This guarantees that there exist M column vectors $u_1,\ldots,u_M \in \{0,1\}^s$ forming a binary CA U of strength $k-1$, and that there exist at least n distinct subsets $Q_1,\ldots,Q_n \in 2^{[M]}$, each containing a set of column indices. Let us access the i-th entry of the column vector u_j via functional notation $u_j(i)$. The trick is now to introduce s linear orders $<_i$, $i=1,\ldots,s$, on $[n]$, which rely on these subsets and are defined as $(a \neq b)$

$$a <_i b :\Leftrightarrow \left(u_{j(a,b)}(i) = 1 \wedge j(a,b) \in Q_a\right) \vee \left(u_{j(a,b)}(i) = 0 \wedge j(a,b) \in Q_b\right). \tag{4.9}$$

[1] The naming appears e.g. in [LKL+11].

[2] Indeed, $2^k\binom{n}{k}(1-\frac{1}{2^k})^d < 2^k\left(\frac{ne}{k}\right)^k(1-\frac{1}{2^k})^d < 2^k\left(\frac{ne}{k}\right)^k\left(e^{-2^{-k}}\right)^d$.

Hereby, $j(a,b) := \min(Q_a \Delta Q_b)$, with Δ denoting the symmetric difference.

We claim now that jointly these orders are k-scrambling (the orders possess equivalent scrambling permutations, see Remark 2.2.10). Let $(a, b_1, \ldots, b_{k-1}) \in S_{n,k}$. Among the defined orders we have to spot one with respect to which all sequences $(a, b_1), \ldots, (a, b_{k-1}) \in [n]^2$ are increasing. Define $P := \{\min(Q_a \Delta Q_{b_\ell}) : \ell = 1, \ldots, k-1\} \subseteq [M]$, having cardinality $|P| \leq k-1$. Let P be enlisted by the sequence of numbers $p_1 < \ldots < p_{|P|}$. Build now the binary word $w \in \{0,1\}^{|P|}$ with letters defined as

$$w_j := \begin{cases} 1, & p_j \in Q_a \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, \ldots, |P|. \tag{4.10}$$

The sought order is now chosen as $<_i$, where i is one of the (for sure existing) row indices of U having $(u_{p_1}(i), \ldots, u_{p_{|P|}(i)}) = w$.

In a case distinction, we assure ourselves that $<_i$ meets the desired requirements. Let $b \in \{b_1, \ldots, b_{k-1}\}$ and consider $j(a,b) = \min(Q_a \Delta Q_b)$ corresponding to, say, the ℓ-th member of the chain $p_1 < \ldots < p_{|P|}$.

Case $j(a,b) \in Q_a$: We observe that also $p_\ell \in Q_a$, and therefore $u_{j(a,b)}(i) = u_{p_\ell}(i) = w_\ell = 1$. We obtain $a <_i b$.

Case $j(a,b) \notin Q_a$: The equalities $u_{j(a,b)}(i) = u_{p_\ell}(i) = w_\ell = 0$ apply. We also necessarily have $j(a,b) \in Q_b$ and therefore conclude $a <_i b$.

It remains to estimate s by n and k. By the minimality in (4.8) and by the previous remark (see (4.7)),

$$\log_2 n > \mathrm{CAK}_{\mathrm{bin}}(s-1, k-1) \geq \frac{k-1}{2e} e^{(s-1)2^{-k+1}(k-1)^{-1}} - 1. \tag{4.11}$$

Solving for s, finally shows $s \leq k2^k(1 + o_n(1)) \log_2 \log_2 n$, which concludes the proof. □

Special cases and optimizations

In [MS03] a slight improvement of the lower bound (4.1) in Theorem 4.1.1 is pointed out (stated below as Theorem 4.1.7). The following basic observation is employed for that purpose. It permits to compare (by iterated substitution) $N(n,k)$ to $N(n,3)$ in a more accurate manner than in (4.1)-(4.2).

Lemma 4.1.6 ([MS03])**.** *For $n \geq k \geq 2$ we can conclude $N(n,k) \geq N(n-1,k-1)+1$.* □

Theorem 4.1.7 ([MS03])**.** *If $n \geq k \geq 2$, then $N(n,k) \geq \log_2 \log_2 (n-k+2) + k - 2$.* □

After introducing k-scrambling permutations, Dushnik provided an explicit formula (see (4.12)), for $N(n,k)$ for large values of k compared to n, i.e., $k \geq 2\sqrt{n} - 2$. For application

purposes, we will see that it is, however, of higher relevance when k is a (small) fixed value.

Theorem 4.1.8 ([Dus50])**.** *Let $k \in \mathbb{N}^{\times}$ be a given strength, and let $\alpha \in \{1, 2\}$. If the permutation length n is parametrized as $n = k^2 + \alpha k$, then*

$$N(n-1, 2k+\alpha-2) = N(n, 2k+\alpha-2) = n - k. \tag{4.12}$$

□

For the special case of $k = 3$, Hajnal noticed that further reduction of the bound can be obtained: The explicitly known number $\mathrm{CAN}_{\mathrm{bin}}(n, 2)$ (derivable via the Erdős-Ko-Rado theorem proved in [EKR61], cf. [KS73]) can be used in the proof of Theorem 4.1.1, on page 29. In [Raj18] it is recognized that a result in [HM99] leads to a lower bound being tight with respect to the upper bound of Hajnal for $k = 3$. We state the bounds together, they yield an explicit expression for $N(n, 3)$ for almost all n.

Theorem 4.1.9. *We have $b(n) + o_n(1) \leq N(n, 3) \leq b(n) + 1 + o_n(1)$, where*

$$b(n) := \log_2 \log_2 n + \frac{1}{2} \log_2 \log_2 \log_2 n + \frac{1}{2} \log_2 \pi. \tag{4.13}$$

□

4.2 Completely k-scrambling permutations

We discuss upper and lower bounds. It will turn out that the additional requirement of completeness enforces logarithmic (upper and lower) bounds in contrast to $\log_2 \log_2$-bounds of the previous section. We explain how to quickly find a logarithmic upper bound using a probabilistic argument due to Spencer [Spe71]. Afterwards, for the derivation of logarithmic lower bounds, we discuss a method first employed by Füredi [Für96] to settle the case $k = 3$ via entropy methods for extremal combinatorics. We then depict how this method was perfected by Radhakrishnan [Rad03] to settle the general case $k \geq 3$.

4.2.1 Upper bounds

We first discuss how to obtain an upper bound for general k and compare it with a sharpened bound for the special case $k = 3$.

We start with the following basic observation of Spencer [Spe71] which was stated in the context of an analysis families of linear orders. For general k, we will later see that only minor improvements are obtainable by constructive approaches ensuring the maintenance of this probabilistic bound, cf. Chapter 5).

Proposition 4.2.1. *Let $n \geq k \geq 3$. The following upper bound applies (k fixed, $n \to \infty$).*

$$N^*(n, k) \leq \frac{k}{\log_2 \frac{k!}{k!-1}} \log_2 n + 1 = \frac{k}{\log_2 \frac{k!}{k!-1}} (1 + o_n(1)) \cdot \log_2 n. \tag{4.14}$$

Proof. For $(p_1, \ldots, p_k) \in S_{n,k}$, we have $|\mathcal{ASC}_{(p_1,\ldots,p_k)}| = n!/k!$. Among all permutations in S_n, the proportion of those ones not appertaining to $\mathcal{ASC}_{(p_1,\ldots,p_k)}$ is therefore $1 - \frac{1}{k!}$.

In total there are $\binom{n}{k}k!$ position selections $(p_1, \ldots, p_k) \in S_{n,k}$ for which potentially there exists no $i \in [d]$ such that $\pi_i(p_1) < \ldots < \pi_i(p_k)$.

Let now $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$ contain permutations which are randomly sampled from S_n (independently and with uniform probability).

For a fixed $(p_1, \ldots, p_k) \in S_{n,k}$, the probability that none of the elements in $\mathscr{P}$ is contained in $\mathcal{ASC}_{(p_1,\ldots,p_k)}$ is given by $(1 - \frac{1}{k!})^d$. Now, whenever

$$\begin{aligned} 1 &> \sum_{(p_1,\ldots,p_k)\in S_{n,k}} \Pr\left[\text{For each } \pi \in \mathscr{F}\colon\ \pi \notin \mathcal{ASC}_{(p_1,\ldots,p_k)}\right] \\ &\geq \Pr\left[\bigvee_{(p_1,\ldots,p_k)\in S_{n,k}} [\text{For each } \pi \in \mathscr{F}\colon\ \pi \notin \mathcal{ASC}_{(p_1,\ldots,p_k)}]\right], \end{aligned}$$

we encounter for sure (positive counter-probability) the possibility that a sampled family violates no monotonicity requirement. This occurs when d is so large that

$$\binom{n}{k}k!(1 - \frac{1}{k!})^d < 1, \tag{4.15}$$

meaning that $d > \frac{\log_2 (n!/(n-k)!)}{\log_2 (k!/(k!-1))}$. Consequently,

$$N^*(n,k) \leq \left\lceil \frac{\log_2 \frac{n!}{(n-k)!}}{\log_2 \frac{k!}{k!-1}} \right\rceil \leq \frac{\log_2 \frac{n!}{(n-k)!}}{\log_2 \frac{k!}{k!-1}} + 1 < \frac{k}{\log_2 \frac{k!}{k!-1}} \log_2 n + 1.$$

□

Remark 4.2.2. *The existence result in Proposition 4.2.1 is non-constructive and leaves open how to actually find a family of permutations testifying the bound.*

For the special case of $k = 3$, the bound (4.14) is asymptotically equal to $C \log_2 n$ with $C \approx 11.405$. Tarui, in [Tar08], due to his explicit construction of completely 3-scrambling permutations, improved considerably the coefficient of the logarithm to $C = 2$. We state this improvement below (more details are discussed in Theorem 5.2.1 and Algorithm 2 of Chapter 5).

Theorem 4.2.3 ([Tar08])**.** *For $n \geq 3$, we have*

$$N^*(n,3) \leq 2\log_2 n + (1 + o_n(1)) \log_2 \log_2 n.$$

□

4.2.2 Lower bounds for strength three

In this section we explain how a lower bound for $N^*(n,3)$ can be obtained by an approach due to Füredi [Für96]. His first key observation is that a result of [KSS81] regarding the entropy of hypergraphs is extendable to the more general structure of multihypergraphs. Moreover, Füredi's original proof was addressed to simultaneously solve a problem in combinatorial geometry closely related to the estimation of $N^*(n,3)$ (cf. Chapter 7 and cf. also [Ish95]).

Definition 4.2.4 (Multihypergraph)**.** *A pair $(V,\mathcal{F})$, where V is a finite set of* vertices *and $\mathcal{F} = \left\{F_1,\dots,F_{|\mathcal{F}|}\right\}$ is a finite multiset of subsets of V with specified multiplicities $\nu_1,\dots,\nu_{|\mathcal{F}|} \in \mathbb{N}^\times$, is called* multihypergraph*. Each set $F_i \in \mathcal{F}$ is called multihyperedge. Set $\|\mathcal{F}\| := \sum_{i=1}^{|\mathcal{F}|} \nu_i$ and call this quantity the weight of $\mathcal{F}$.*

Let us state and prove the aforementioned result regarding the entropy of multihypergraphs.

Theorem 4.2.5 ([Für96], cf. also [KSS81])**.** *Consider a multihypergraph $(V,\mathcal{F})$ whose multihyperedges $F_1,\dots,F_{|\mathcal{F}|}$ have multiplicities $\nu_1,\dots,\nu_{|\mathcal{F}|}$ and maximum multiplicity $\mu_{\max} = \max_{i=1,\dots,|\mathcal{F}|}\nu_i$. For $v \in V$ denote by*

$$\alpha_v := \frac{1}{\|\mathcal{F}\|}\sum_{i=1}^{|\mathcal{F}|} \mathbb{1}_{F_i}(v)\nu_i$$

the proportion (weighted according to multiplicities) of elements in $\mathcal{F}$ containing v. If S is a random variable attaining values in $\mathcal{F}$ and has associated probability mass function $p_S(F_i) = \frac{\nu_i}{\|\mathcal{F}\|}$, then

$$\log_2\left(\frac{\|\mathcal{F}\|}{\mu_{\max}}\right) \le H[S] \le \sum_{v\in V} \mathsf{H}(\alpha_v)\,. \tag{4.16}$$

Proof. Noticing

$$\log_2\left(\frac{\|\mathcal{F}\|}{\mu_{\max}}\right) = \sum_{i=1}^{|\mathcal{F}|}\frac{\nu_i}{\|\mathcal{F}\|}\log_2\left(\frac{\|\mathcal{F}\|}{\mu_{\max}}\right) \tag{4.17}$$

$$\le \sum_{i=1}^{|\mathcal{F}|}\frac{\nu_i}{\|\mathcal{F}\|}\log_2\left(\frac{\|\mathcal{F}\|}{\nu_i}\right) = H[S], \tag{4.18}$$

the left inequality is shown.

For the right inequality we proceed as follows. Identify V without loss of generality with $\{1,\dots,|V|\} \subseteq \mathbb{N}$. Then, an element $F \in \mathcal{F} \subseteq 2^{\{1,\dots,|V|\}}$ can be encoded by an indicator tuple whose entries attain values in $\{0,1\}$: The i-th entry attains the value 1 iff $i \in F$, $i \in \{1,\dots,|V|\}$. As the encoding is bijective the entropy remains invariant under the

proposed transformation (we assume the probability mass function is inherited via the bijection).

Let S now be a random variable attaining values in $\mathcal{F}$ with probability mass function $p_S(F_i) = \frac{\nu_i}{\|\mathcal{F}\|}$. Furthermore, consider the random variables $C_i(S)$ attaining values in $\{0,1\}$, $i = 1, \ldots, |V|$, that return 1 iff $i \in S$. The entropy of S can be expressed as joint entropy, i.e., $H(S) = H[C_1(S), \ldots, C_{|V|}(S)]$ (recall the previous bijective conversion to an indicator tuple). By the independence bound (see Corollary 3.2.8), the joint entropy can be bounded by

$$H[C_1(S), \ldots, C_{|V|}(S)] \leq \sum_{i=1}^{|V|} H[C_i(S)] = \sum_{i=1}^{|V|} \mathsf{H}(\alpha_i)\,. \tag{4.19}$$

□

We need to introduce a few auxiliary sets/quantities.

Definition 4.2.6. *For a given d-family $\mathscr{P} \subseteq S_n$ and fixed $\varepsilon \in \{-1,1\}^d$, set*

$$L(v, \varepsilon, \mathscr{P}) := \{w \in [n] \setminus \{v\} : \varepsilon_i = -1 \Leftrightarrow \pi_i(w) < \pi_i(v) \text{ for all } i = 1, \ldots, d\}\,, \tag{4.20}$$

$$\ell(\mathscr{P}) := \max\left\{|L(v, \varepsilon, \mathscr{P})| : (v, \varepsilon) \in [n] \times \{-1,1\}^d\right\}, \tag{4.21}$$

and

$$\ell(n, d) := \min\{\ell(\mathscr{P}) : \mathscr{P} \subseteq S_n \text{ has cardinality } d\}\,. \tag{4.22}$$

Lemma 4.2.7. *Fix $d \in \mathbb{N}^\times$. If there exists a 3-mixing family $\mathscr{F} \subseteq S_n$ consisting of d permutations, then necessarily $\ell(n, d) \leq 1$.*

Proof. Let $\mathscr{P}$ be a 3-mixing family of d permutations of $[n]$. Seeking a contradiction, assume now $\ell(n, d) \geq 2$. In particular, $\ell(\mathscr{P}) \geq 2$ and therefore for $\mathscr{P}$ the following must be true: There exist $v \in [n]$ and $\varepsilon \in \{-1,-1\}^d$ such that $L(v, \varepsilon, \mathscr{P})$ contains two different elements $w_1, w_2 \in [n] \setminus \{v\}$.

We consider now the position triple (v, w_1, w_2). As $\mathscr{P}$ is 3-mixing, we can find an index $i \in [d]$ such that without loss of generality

$$\pi_i(w_1) < \pi_i(v) < \pi_i(w_2).$$

The latter chain always contradicts the property

$$\operatorname{sgn}(\pi_i(w_1) - \pi_i(v)) = \operatorname{sgn}(\varepsilon_i) = \operatorname{sgn}(\pi_i(w_2) - \pi_i(v))\,,$$

which, however, must hold due to the appertainance $w_1, w_2 \in L(v, \varepsilon, \mathscr{P})$. □

Theorem 4.2.8 ([Für96, Theorem 1.2])**.** *For $n, d \geq 3$ we have*

$$\ell(n, d) > (n-1)e^{-\frac{d}{2}\frac{n-1}{n}}\,. \tag{4.23}$$

Proof. Assume $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$. Furthermore, suppose $\mathscr{P}$ is a minimizer of $\ell(\cdot,\cdot)$ in (4.22), i.e., $\ell(n,d) = \ell(\mathscr{P})$. Given $x, y \in [n]$, consider their induced subset $F(y,x) \subseteq [d]$ given by

$$F(y,x) := \{i \in [d] : \pi_i(y) < \pi_i(x)\}\,. \tag{4.24}$$

Fix $x \in [n]$ and consider the multihypergraph $(V, \mathcal{F}(x))$ formed by $V := [n]$ and

$$\mathcal{F}(x) := \{F(y,x) : y \in [n] \setminus \{x\}\}\,, \tag{4.25}$$

where we keep the multiplicities for the latter system of sets. Let $s = |\mathcal{F}(x)|$ and let us denote by $B_1, \ldots, B_s$ the pairwise distinct sets composing $\mathcal{F}(x)$ (with corresponding multiplicities $\nu_1, \ldots, \nu_s$ satisfying $\sum_{i=1}^{s} \nu_j = \|\mathcal{F}(x)\| = n-1$).

We remark that no multiplicity ν_j, $j = 1, \ldots, s$, can exceed the upper bound $\ell(\mathscr{P})$: Fix ν_j elements $z_1, \ldots, z_{\nu_j}$ satisfying $B_j = F(z_1, x) = \ldots = F(z_{\nu_j}, x)$. For $i \in [d] \setminus B_j$ we have $\pi_i(z_r) > \pi_i(x)$, $r = 1, \ldots \nu_j$. Define $\varepsilon \in \{-1,1\}^d$ with $\varepsilon_i = -1$ iff $i \in \left\{z_1, \ldots, z_{\nu_j}\right\}$. Then, ν_j corresponds to $|L(x, \varepsilon, \mathscr{P})|$ which cannot exceed $\ell(\mathscr{P}) = \max\left\{|L(v, \varepsilon, \mathscr{P})| : (v,\varepsilon) \in [n] \times \{-1,1\}^d\right\}$. By arbitrariness of ν_j we have additionally shown $\mu_{\max}(\mathcal{F}(x)) \leq \ell(\mathscr{P})$.

Using the inequality concerning the entropy of a multihypergraph, see Theorem 4.2.5, we can estimate (we notice that the element i appears in the members of $\mathcal{F}(x)$ exactly $\pi_i(x) - 1$ times)

$$\log_2\left(\frac{n-1}{\ell(\mathscr{P})}\right) \leq \log_2\left(\frac{n-1}{\mu_{\max}(\mathcal{F}(x))}\right) \leq \sum_{i=1}^{d} \mathsf{H}\Big(\frac{\pi_i(x)-1}{n-1}\Big)\,. \tag{4.26}$$

By arbitrariness of x, this applies also when averaging over $x \in [n]$, i.e.,

$$\log_2\left(\frac{n-1}{\ell(\mathscr{P})}\right) \leq \frac{1}{n}\sum_{x=1}^{n}\sum_{i=1}^{d} \mathsf{H}\Big(\frac{\pi_i(x)-1}{n-1}\Big) = \frac{d}{n}\sum_{j=0}^{n-1} \mathsf{H}\Big(\frac{j}{n-1}\Big) \tag{4.27}$$

$$< (n-1)\frac{d}{n}\int_0^1 \mathsf{H}(\xi)\,\mathrm{d}\,\xi \tag{4.28}$$

$$= -d\frac{n-1}{n}\int_0^1 (\xi \log_2(\xi) + (1-\xi)\log_2(1-\xi))\,\mathrm{d}\,\xi \tag{4.29}$$

$$= d\frac{n-1}{n}\frac{\log_2(e)}{2}. \tag{4.30}$$

Hereby (4.28) follows from concavity of $\mathsf{H}(\xi)$ and the fact that $\xi = \frac{1}{2}$ is its symmetry axis. Comparison of first and last member in (4.27)-(4.30) yields

$$\ell(\mathscr{P}) > (n-1)e^{-\frac{d}{2}\frac{n-1}{n}}.$$

□

A combination of the previous results shows the –according to the state of the art– best lower bound for completely 3-scrambling permutations.

Corollary 4.2.9 (Füredi's lower bound)**.** *For all $n \geq 4$, we have*

$$2\ln(2)\log_2 n + 1 < N^*(n,3)\,. \tag{4.31}$$

Proof. Consider a fixed $n \geq 3$. Choose $d \in \mathbb{N}^\times$ large enough such that a 3-mixing d-family $\mathscr{P}$ of permutations of $[n]$ exists. From (4.23) combined with Lemma 4.2.7 we obtain that $(n-1)e^{-\frac{d}{2}\frac{n-1}{n}} < 1$ allowing to conclude $d > 2\ln(2)\frac{n}{n-1}\log_2(n-1)$. Provided $n \geq 4$, we therefore have

$$2\ln(2)\log_2 n < 2\ln(2)\frac{n}{n-1}\log_2(n-1) \leq N^{\mathrm{mix}}(n,3) \leq N^*(n,3) - 1. \tag{4.32}$$

The weakened bound on the left of (4.32) can be used as simpler bound. The relation between N^{mix} and N^* was already observed in Remark 2.2.5. □

Remark 4.2.10. *In [Rad03] it is claimed that for $k = 3$ Füredi's bound is improved by a factor of "approximately $2/\log_2(e)$". However, we have discussed Füredi's proof in great detail and in some parts renounced to use generous estimates (being applied in the original work); we arrived at the same bound as in [Rad03] (consult the bound for $k = 3$ in the subsequent Theorem 4.2.15).*

4.2.3 Lower bounds for arbitrary strength

We will fall back on the Fredman-Komlós bound. It first appears in [FK84] where the authors studied colorings of graphs with colors being d dimensional vectors over a finite alphabet including a wildcard symbol. They examined a particular subclass of those colorings leading to their notion of the *content of a bipartite graph* and a lower bound for it. Below we make use of a simpler proof strategy due to [Rad01] allowing furthermore to relax the requirement of being bipartite.[3]

We recall that a *proper coloring* of a graph is a mapping from its vertex set to a finite set of colors such that each two adjacent vertices are mapped to different colors – a partition (in whose entropy one is interested) of the vertices in color classes is hereby induced.

Definition 4.2.11 (Content of a graph, [Rad01, FK84])**.** *Let $G = (V, E)$ be a graph. Denote by V^{isol} the subset of* isolated[4] *vertices and by $V^{nonisol}$ its complementary vertices. Let $\hat{\chi}$ be a proper coloring of the vertices in $V^{nonisol}$ such that $H[Z]$ is minimal (where $Z \in V^{nonisol}$ is a vertex drawn uniformly at random). Define as the* content *of G the quantity*

$$\mathrm{content}(G) := \tfrac{|V^{nonisol}|}{|V|} H[\hat{\chi}(Z)]. \tag{4.33}$$

[3] Moreover, this bound is a weaker form of the closely related subadditivity property of the so-called (Körner) entropy of a graph (extensive discussion of graph entropy (including properties, equivalent formulations, etc.) can be found in [Kör73, Rad01]).

[4] A vertex is isolated if it does not possess any incident edges.

Lemma 4.2.12 (Fredman-Komlós bound [Rad01, FK84]). *Let $G = (V, E)$, $G_1 = (V, E_1)$,...,$G_t = (V, E_t)$ be undirected graphs and assume $\bigcup_{i=1}^{t} G_i = G$ (in terms of edge sets). Then, with $\alpha(G)$ denoting the maximum size of an independent set of G, we have*

$$\sum_{i=1}^{t} \mathrm{content}(G_i) \geq \log_2 \left(\frac{n}{\alpha(G)} \right). \tag{4.34}$$

Proof. Consider for each G_i a fixed proper coloring $\chi_i : V \to \mathbb{N}$. The key idea is to use the following fact: For an unknown vertex $X \in [n]$, knowing inside each G_i only the color class of X disturbed by some "noise" arising from the isolated vertices in V_i^{isol} still permits to recover valuable information about how many potential values X might have attained. In fact, the cardinality of

$$M = \left(\chi_1^{-1}(\{\chi_1(X)\}) \cup V_1^{\mathrm{isol}} \right) \cap \ldots \cap \left(\chi_t^{-1}(\{\chi_t(X)\}) \cup V_t^{\mathrm{isol}} \right) \supseteq \{X\} \tag{4.35}$$

cannot exceed $\alpha(G)$: If there was a subset of $\alpha(G) + 1$ vertices being contained in each intersectand of M (each intersectand being the union of a color class with a set of isolated points is automatically an independent set) this would together with the covering condition $G = \bigcup G_i$ contradict maximality of $\alpha(G)$.

Assume now X is a (with uniform probability chosen) random vertex of V and that $Z(X)$ is an independently of X randomly chosen vertex of V^{nonisol} (again uniform probability of choice). Define now the random variables

$$Y_i := \begin{cases} \chi_i(X), & X \in V^{\mathrm{nonisol}} \\ \chi_i(Z(X)), & \text{otherwise} \end{cases}, \quad i = 1, \ldots, t.$$

By the previous observation concerning M, we can now bound from above the entropy of X conditioned to $(Y_1, \ldots, Y_t)$ and hereby derive (by the chain rule and the independence bound, cf. Section 3.2)

$$\begin{aligned} \log_2 \alpha(G) &\geq H[X|Y_1, \ldots, Y_t] && (4.36)\\ &= H[X, Y_1, \ldots, Y_t] - H[Y_1, \ldots, Y_t] && (4.37)\\ &= H[X] + H[Y_1, \ldots, Y_t|X] - H[Y_1, \ldots, Y_t] && (4.38)\\ &\geq \log_2 |V| + H[Y_1, \ldots, Y_t|X] - \sum_{i=1}^{t} H[Y_i]. && (4.39) \end{aligned}$$

By construction, $Y_1|X, \ldots, Y_t|X$ are independent random variables. We use this fact to

rewrite the middle summand in (4.39) and conclude

$$\log_2 \alpha(G) \geq \log_2 |V| + \sum_{i=1}^{t} H[Y_i|X] - \sum_{i=1}^{t} H[Y_i] \tag{4.40}$$

$$= \log_2 |V| + \sum_{i=1}^{t} \left(1 - \frac{V_i^{\text{nonisol}}}{|V|}\right) H[Y_i] - \sum_{i=1}^{t} H[Y_i] \tag{4.41}$$

$$= \log_2 |V| - \sum_{i=1}^{t} \frac{V_i^{\text{nonisol}}}{|V|} H[Y_i]. \tag{4.42}$$

By Definition 4.2.11 we finally obtain

$$\sum_{i=1}^{t} \text{content}(G_i) \geq \log_2 \frac{|V|}{\alpha(G)}.$$

□

Example 4.2.13. *The complete graph K_n on the vertex set $[n]$ has only singletons as independent sets. Hence, $\log_2 n$ is the respective lower bound in* (4.34) *for K_n.*

For a tuple $q \in S_{n,k-2}$, we recall the notation

$$\mathcal{ASC}_q^{[k-2]} := \{\pi : \pi \in S_n \text{ with } \pi(q_1) < \pi(q_2) < \ldots < \pi(q_{k-2})\}, \tag{4.43}$$

which contains a fraction of $\frac{1}{(k-2)!}$ permutations of the entire group S_n. We provide in advance the following auxiliary estimate resulting from a calculation in [Rad03] (its proof is suppressed an depends as in the previously handled case $k = 3$ on the symmetry and concavity of the binary entropy function).

Lemma 4.2.14 ([Rad03])**.** *For each combination of a permutation $\pi \in S_n$ and a sequence $q \in S_{n,k-2}$, consider the graph $G_q(\pi)$ possessing the set of vertices $V := [n] \setminus \{q_1, \ldots, q_{k-2}\}$ and the set of edges*

$$E = \{\{i, j\} : \pi(i) < \pi(q_1) < \pi(q_2) < \ldots < \pi(q_{k-2}) < \pi_j\}. \tag{4.44}$$

If $\pi \in S_n$ is fixed, then, for the conditional expectation, we have

$$\mathop{\mathbb{E}}_{q \in S_{n,k-2}} \left[\text{content}(G_q(\pi)) \,|\, \pi \in \mathcal{ASC}_q^{[k-2]}\right] \leq \frac{\log_2 e}{2} \frac{2n - k + 1}{n(k-1)}. \tag{4.45}$$

□

The latter result is useful for obtaining the following estimate.

Theorem 4.2.15 ([Rad03])**.** *Let $n > k \geq 3$. We can estimate*

$$N^*(n,k) > \frac{2}{\log_2 e}\frac{n(k-1)!}{2n-k+1}\log_2\,(n-k+2)+1, \tag{4.46}$$

such that, for fixed k and $n \to \infty$,

$$N^*(n,k) \geq \frac{(k-1)!}{\log_2 e}(1+o_n(1))\log_2 n. \tag{4.47}$$

Proof. We use the fact that $N^*(n,k)-1 \geq N^{\mathrm{mix}}(n,k)$ and bound from below the latter quantity. Let $\mathscr{F}$ be a k-mixing d-family of permutations of $[n]$. We make use of the graph $G_q(\pi)$ of Lemma 4.2.14 which is bipartite and potentially has empty edge set (in case $\pi \notin \mathcal{ASC}_q^{[k-2]}$). For arbitrary but fixed $q \in S_{n,k-2}$, the union over all $\pi \in \mathscr{F}$ of all $G_q(\pi)$ is (by k-mixingness) equal to the complete graph on the vertex set V.

When $\pi \in \mathcal{ASC}_q^{[k-2]}$, then $\pi(q_1)-1$ (respectively $n-\pi(q_2)$) quantifies the number of vertices being non-isolated and contained in the "left" (respectively "right") side of the bipartite graph (see (4.44)). Consequently, the content of such a bipartite graph given by

$$\mathrm{content}(G_q(\pi)) = \frac{n-\pi(q_{k-2})+\pi(q_1)-1}{n-k+2}\;\mathsf{H}\Big(\frac{\pi_{q_1}-1}{n-\pi(q_{k-2})+\pi(q_1)-1}\Big),$$

where we apply the convention that the second factor vanishes if its argument's denominator does.

The prerequisites for Fredman-Komlós estimate (4.34) are satisfied and hence, together with the arbitrariness of $q \in S_{n,k-2}$, the estimate implies

$$\sum_{\pi\in\mathscr{F}}\mathop{\mathbb{E}}_{q\in S_{n,k-2}}[\mathrm{content}(G_q(\pi))] = \mathop{\mathbb{E}}_{q\in S_{n,k-2}}\left[\sum_{\pi\in\mathscr{F}}\mathrm{content}(G_q(\pi))\right] \geq \log_2\,(n-k+2)\,. \tag{4.48}$$

Hereby, all permutations $\tilde{q}$ satisfying $\pi \notin \mathcal{ASC}_{\tilde{q}}^{[k-2]}$ do not contribute to the members of the left sum in (4.48). It can therefore be rewritten and estimated with help of Lemma 4.2.14:

$$\begin{aligned}\sum_{\pi\in\mathscr{F}}\mathop{\mathbb{E}}_{q\in S_{n,k-2}}[\mathrm{content}(G_q(\pi))] &= \sum_{\pi\in\mathscr{F}}\frac{1}{(k-2)!}\mathop{\mathbb{E}}_{q\in S_{n,k-2}}\Big[\mathrm{content}(G_q(\pi))\,|\,\pi\in\mathcal{ASC}_q^{[k-2]}\Big]\\ &\leq \sum_{\pi\in\mathscr{F}}\frac{1}{(k-2)!}\frac{\log_2 e}{2}\frac{2n-k+1}{n(k-1)}\\ &= d\frac{1}{(k-2)!}\frac{\log_2 e}{2}\frac{2n-k+1}{n(k-1)}.\end{aligned} \tag{4.49}$$

The assertion ensues, as (4.49) bounds (4.48) from above. □

4.2.4 An alternative lower bound for arbitrary k

We show how to derive an alternative lower bound for the cardinality of completely k-scrambling families of permutations (or equivalently for the cardinality of sequence covering arrays of strength k). We will show, however, that this bound cannot compete with the one achieved by the approach of Füredi-Radhakrishnan, cf. (4.47). For this purpose, we adapt a "symbiosis" of SCAs and binary CAs pointed out in [KHL+12] for strength three and generalize it to arbitrary strength k.

Let us denote the asymptotic coefficient (independent of n) of the logarithmic term in (4.47) by $c_{\mathrm{FR}}(k) := (k-1)!/\log_2 e$. We will derive a different lower bound asymptotically corresponding as well to a constant multiple of the logarithm (for $n \to \infty$ and k fixed). Afterwards we analyze the quality of this alternative coefficient by comparing it with $c_{\mathrm{FR}}(k)$.

Theorem 4.2.16. *Let $k \geq 4$ and let $\vartheta(k) := (k-2)\left(\mathsf{H}\left(2^{1-k}\right) - 2^{2-k}\right)^{-1}$ with $\mathsf{H}(\cdot)$ denoting the binary entropy function. Then, there is a minorant of $N^*(n,k)$ having asymptotic growth $c_{\mathrm{KS}}(k)\log_2 n$ where $c_{\mathrm{KS}}(k) := \vartheta(k-1)$.*

To prove the latter theorem, the following result is essential. It states that a completely k-scrambling d-family allows to constructively derive from it a binary covering array of strength $k-1$ on d rows.

Lemma 4.2.17. *Let $n \geq k \geq 3$ and let $\mathscr{P} = \{\pi_1, \ldots, \pi_d\} \subseteq S_n$ be a completely k-scrambling family. Then, the matrix $B = (b_{ij})_{i,j} \in \{0,1\}^{d\times(n-1)}$ with entries*

$$b_{i,j-1} := \begin{cases} 1, & \pi_i(1) < \pi_i(j) \\ 0, & \text{otherwise} \end{cases}, \quad j = 2, \ldots, n, \tag{4.50}$$

is a $d \times (n-1)$ covering array over the alphabet $\{0,1\}$ with strength parameter $k-1$.

Proof. Consider a $(k-1)$-tuple of increasing indices $1 \leq j_1 < \ldots < j_{k-1} \leq n-1$. We aim to show that it is possible to spot an arbitrary binary word $w \in \{0,1\}^{k-1}$ among the rows of the submatrix of B consisting of the d shrunken rows $(b_{i,j_\ell})_{\ell=1}^{k-1}$, $i = 1, \ldots, d$.

Let m be the count of zeros in w. Denote by z_j, $j = 1, \ldots, m$, the position (within) w of the j-th zero. Similarly, the positions of the ones are enlisted as u_j, $j = 1, \ldots, k-1-m$. This means that

$$\{z_j : j \in [m]\} \cup \{u_j : j \in [k-1-m]\} = [k-1].$$

For $\pi \in \mathscr{P}$, consider now the $((z_j), 1, (u_j))$-induced "traversal" of π, i.e., the sequence

$$\pi(j_{z_1}), \pi(j_{z_2}), \ldots, \pi(j_{z_m}), \pi(1), \pi(j_{u_1}), \pi(j_{u_2}), \ldots, \pi(j_{u_{k-1-m}}),$$

which is (as $\mathscr{P}$ is completely k-scrambling) monotonically increasing for suitable $\hat{\pi} \in \mathscr{P}$. Therefore, the row-vector resulting from $\hat{\pi}$ via the construction (4.50) will have zeros, ones at the positions j_{z_ℓ}, j_{u_ℓ}, respectively. It hence corresponds to w. This completes the proof. □

Directly from Lemma 4.2.17 we obtain the following result.

Theorem 4.2.18. *For $n \geq k \geq 2$, we have $N^*(n,k) \geq \mathrm{CAN}_{\mathrm{bin}}(n, k-1)$.* □

Kleitman and Spencer have obtained the subsequent lower bound for binary covering arrays.

Theorem 4.2.19 ([KS73])**.** *With fixed $k \geq 3$ and $\vartheta(k)$ defined as in Theorem 4.2.16 we have (k fixed, $n \to \infty$)*

$$\mathrm{CAN}_{\mathrm{bin}}(n,k) \geq (\vartheta(k) + o_n(1)) \log_2 n. \tag{4.51}$$

□

Finally, we obtain the sought-after assertion.

Proof of Theorem 4.2.16. Combining Theorem 4.2.18 with Theorem 4.2.19 yields the claim. □

In Figure 4.1 we compare the values of c_{FR} and c_{KS}: While the trend for large k is clear, the plot attests better performance to c_{FR} as well for small values of k. In particular, we get a numeric impression of how costly even tiny increases in k impact on the size of the permutation families.

Also for the case $k = 3$, which is not included in Theorem 4.2.16, we cannot obtain an improvement to Füredi's lower bound (4.2.9) via the above argumentation. To see this, we refer to the fact that the $N^*(n,3)$-minorizing quantity $\mathrm{CAN}_{\mathrm{bin}}(n,2)$ can even be determined explicitly (as already mentioned in Section 4.1). The coefficient of the (dominating) logarithmic term is hereby exactly one, being less than Füredi's asymptotic coefficient ($c_{\mathrm{FR}}(3) \approx 1.386$).

In the following sections, we turn our attention to the regularized forms of the discussed concepts.

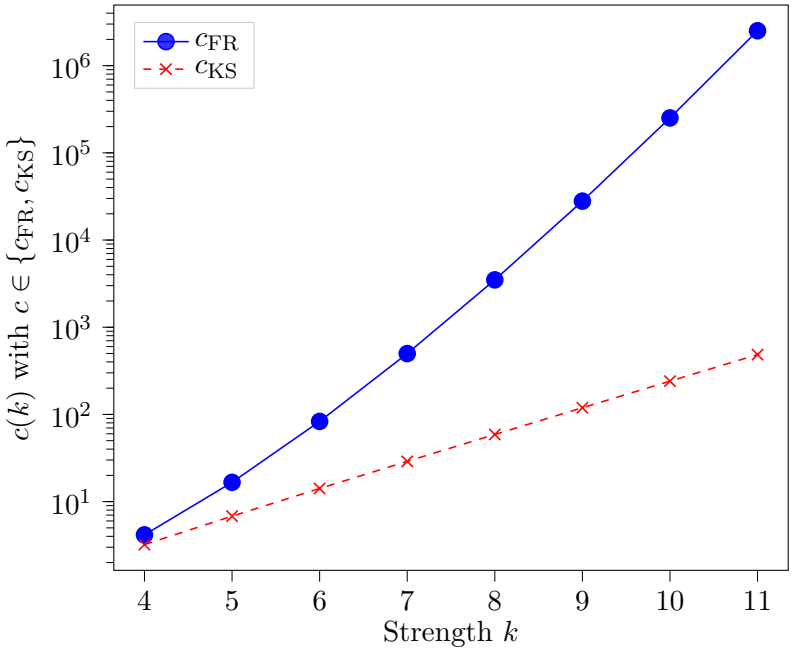

Figure 4.1: A comparison of lower bound coefficients $c(k) \in \{c_{\mathrm{FR}}(k), c_{\mathrm{KS}}(k)\}$ (for small values of k) asymptotically satisfying $N^*(n,k) \geq c(k) \log_2 n$.

4.3 k-restricted min-wise independent permutations

Due to the fact that k-restricted min-wise independent permutations are $\{j\}$-restricted min-wise independent (recall Definition 2.3.15), for $j = 1, \ldots, k$, it is elementary that the bound $|\mathscr{F}| \geq \operatorname{lcm}\{k, k-1, \ldots, 2, 1\} = e^{k-o(k)}$ applies (cf. [BCFM00]). An alternative bound from [ITT00] can be employed attesting that $|\mathscr{F}| \geq n-1$ – depending hence (exclusively) on n and obtained from consideration of the special case $k = 3$. We state a more recent bound improving on the latter two bounds.

Theorem 4.3.1 ([ITT03])**.** *For a k-restricted min-wise independent family $\mathscr{F} \subseteq S_n$, we have $|\mathscr{F}| \geq \binom{n-1}{\lfloor k/2 \rfloor}(1 + o_n(1)) \geq \frac{1+o_n(1)}{\lfloor k/2 \rfloor^{\lfloor k/2 \rfloor}} n^{\lfloor k/2 \rfloor}$, more precisely*

$$|\mathscr{F}| \geq \begin{cases} \sum_{i=0}^{h} \binom{n-1}{i}, & \text{if } k = 2h+1 \quad (h \in \mathbb{N}) \\ \sum_{i=0}^{h} \binom{n-1}{i} + \binom{n-2}{h}, & \text{else, if } k = 2h \quad (h \in \mathbb{N}) \end{cases}.$$

□

Theorem 4.3.2 ([ITT00, Theorem 3.3])**.** *There exists a k-restricted min-wise independent family $\mathscr{F} \subseteq S_n$ which is of size polynomial in n and satisfies more precisely*

$$|\mathscr{F}| \leq 2^k ((k-1)!)^k \, n^k.$$

□

For the approximate concept we obtain the following analogous result.

Theorem 4.3.3 ([BCFM00])**.** *A ε-approximately k-restricted min-wise independent family $\mathscr{F} \subseteq S_n$ of at most cardinality $O(k^2 \ln(n/k)/\varepsilon^2)$ exists. In particular, there exists a ε-approximately min-wise independent family $\mathscr{F} \subseteq S_n$ of cardinality $|\mathscr{F}| \leq O(n^2/\varepsilon^2)$.*

Proof. We will employ the probabilistic method. Consider $\ell \leq k$ and a fixed ℓ-subset $X \subseteq [n]$ with a distinguished element $x \in X$. Suppose we sample (with replacement) d permutations from S_n uniformly at random. For each sampled permutation $\pi \in S_n$, we can check if π is contained in the set $C(X, x)$ containing all permutations $\sigma \in S_n$ satisfying $\sigma(x) = \min \sigma(X)$. We can regard the sampling process as a Bernoulli process with d trials and success probability $p := \Pr[\pi \in C(X, x)] = 1/\ell$. The mean (expected count of successes) is determined by pd. The probability that the output of the process differs from the mean more than ε times the mean can be bounded by the Chernoff bound (3.3) of Section 3.1 as follows:

$$\Pr\left[\left|\sum_{i=1}^{d} \mathbb{1}_{C(X,x)}(\pi) - pd\right| > \varepsilon pd\right] \leq 2e^{-pd\varepsilon^2/3} \leq 2e^{-d/k\varepsilon^2/3}. \tag{4.52}$$

For a constellation (X, x), the probability that at least one choice of (X, x) leads to a relative error larger than εpd in (4.52) is majorized by the union bound

$$\sum_{(X,x)} 2ke^{-d/k\varepsilon^2/3} = \sum_{\ell=1}^{k} \binom{n}{\ell} \ell 2e^{-d/k\varepsilon^2/3} \leq \binom{n+k+1}{k} 2e^{-d/k\varepsilon^2/3}. \tag{4.53}$$

For the choice of the smallest d such that

$$d > \frac{3k}{\varepsilon^2} \ln \binom{n+k+1}{k} + \ln 2, \tag{4.54}$$

the right hand side of (4.53) strictly minorizes 1. This means that after d trials there is the chance that $\left|\sum_{i=1}^{d} \mathbb{1}_{C(X,x)}(\pi) - pd\right| \leq \varepsilon pd$ for any constellation of (X, x). The existence of a ε-approximately k-restricted min-wise independent family of cardinality $d = O(n^2/\ln(n/k))$ is hence granted. □

Remark 4.3.4. *We also tried to use the probabilistic argument in Theorem 4.3.3 for deriving an upper bound for PSCAs/rankwise independent families. The approach was inconclusive. Similarly, we also noticed that other versions of the Chernoff bound estimating the absolute distance from the mean seem not to be helpful for this purpose.*

We now turn to upper bounds, in particular their magnitude for special values like $k = n$ and $k \in \{3, 4\}$.

Itoh et al. [ITT00] have shown that for (n-restricted) min-wise independence the above discussed exponential lower bound $\operatorname{lcm}\{n, n-1, \ldots, 1\}$ is perfectly tight.

Theorem 4.3.5 ([ITT00])**.** *There exists a construction to obtain a min-wise independent family $\mathscr{F} \subseteq S_n$ that has cardinality $|\mathscr{F}| = \operatorname{lcm}\{n, n-1, \ldots, 1\}$.* □

Let us consider the following auxiliary, recursive sequence introduced in [TIT03].

Definition 4.3.6 (L_q)**.** *Let $g : \{2^\tau : \tau \in \mathbb{N}^\times\} \to \mathbb{N}^\times$,*

$$2^\tau = s \mapsto \begin{cases} \sqrt{s}, & \text{if } 2|\tau \\ \sqrt{2s}, & \text{otherwise} \end{cases}. \tag{4.55}$$

Let $q \in \{2^{2\tau} : \tau \in \mathbb{N}^\times\}$ be fixed. Consider the sequence $L_q := (q_\ell, q_{\ell-1}, \ldots, q_2, q_1)$ which is obtained by initially setting $q_\ell := q$, and $q_{\ell-1} = g(q), q_{\ell-2} = g \circ g(q), \ldots, q_1 = 4$ (ℓ is determined implicitly here).

As we will fall back on concrete evaluations of the sequences L_q, for convenience, we prepare its values.

Example 4.3.7 (Lookup table for L_q sequences)**.**

q	τ	L_q
4	1	(4)
16	2	(16, 8, 4)
64	3	(64, 8, 4)
256	4	(256, 16, 4)

Remark 4.3.8. *Tarui et al. [TIT03] come up with a recursive construction for 4-restricted min-wise independent families. We state (a slight improvement of) their result in Theorem 4.3.9 (their leading constant coefficient is easily improvable by a factor of 2 as we show). The cardinalities of the families returned by the recursive construction obey as well a recursion which is homogeneous in the cardinality of the family employed as base case of the recursion (the latter family can be any 4-restricted min-wise independent family $\mathscr{F} \subseteq S_4$). In [TIT03], Tarui et al. choose as base case $\mathscr{F} = S_4$. This can be improved by picking a base case of smaller cardinality (12 permutations are sufficient – this is postponed in Example 4.3.10) and leads to the result stated next.*

Theorem 4.3.9 ([TIT03, Theorem 4] sharpened by factor 2)**.** *Let $n \geq 4$. There exists a 4-restricted min-wise independent family $\mathscr{F} \subseteq S_n$ such that $|\mathscr{F}| \leq 6\sqrt{e}(1+o_n(1)) \cdot n(\log_2 n)^3$.* □

Example 4.3.10. *The general construction provided in [ITT00] and establishing Theorem 4.3.5 allows to construct a (4-restricted) min-wise independent 12-family of permutations in S_4. It is rather involved and so we prefer to construct such a family directly: For this, we take the family of type* PSCA$(4, 3, 1)$ *from Example 2.3.7, call it $\mathscr{A}$, and subject it to the substitution of symbols $\{1 \leftrightarrow 3, 2 \leftrightarrow 4\}$. The result $\tilde{\mathscr{A}} = \{\gamma \circ \pi : \pi \in \mathscr{A}\}$ (with*

$\gamma = (3,4,1,2) \in S_n$) will as well be of type PSCA$(4,3,1)$ *such that $\mathscr{A} \cup \tilde{\mathscr{A}}$ yields a representative of* PSCA$(4,3,2)$. *We can map $\mathscr{A} \cup \tilde{\mathscr{A}}$ via the isomorphism ι of Lemma 2.2.9 to a 3-rankwise independent family of permutations in S_4, which we call $\mathscr{F}$. We state $\mathscr{F}$ explicitly below and notice that, for every $i \in \{1,\dots,4\}$, there are always 3 permutations $\pi \in \mathscr{F}$ for which $\pi(i) = 1$. Combining both observations on $\mathscr{F}$, we therefore obtain that $\mathscr{F}$ is 4-restricted min-wise independent and has cardinality 12.*

$$\begin{aligned}\mathscr{F} = \{\quad & (1,2,4,3), \quad (3,4,1,2),\\ & (2,1,4,3), \quad (4,3,2,1),\\ & (2,4,1,3), \quad (1,3,2,4),\\ & (4,2,1,3), \quad (1,3,4,2),\\ & (2,4,3,1), \quad (3,1,2,4),\\ & (4,2,3,1), \quad (3,1,4,3) \quad \}.\end{aligned}$$

Remark 4.3.11. *The following estimate will be automatically valid also for 3-rankwise independent families (as they are 3-restricted min-wise independent, recall Lemma 2.3.17). It relies again on a recursive construction obtained in [TIT03]. We state it in slightly optimized form exploiting the fact that an argumentation similar to Remark 4.3.8 is applicable also this time (for the base case of recursion $\mathscr{F}$, employ the 3-restricted min-wise independent 6-family $\mathscr{Q}$ of Example 2.2.4 instead of the choice $\mathscr{F} = S_4$ in [TIT03]). Indeed a 3-restricted min-wise independent family $\mathscr{F} \subseteq S_{q_{\ell-1}}$ allows to construct a 3-restricted min-wise independent family $\mathscr{F}' \subseteq S_{q_\ell}$. The accompanying recursion for the cardinalities reads $C(q_\ell) = 2(q_{\ell-1}+1) \cdot C(q_{\ell-1})$.*

Theorem 4.3.12 ([TIT03, Theorem 3] sharpened by factor 4)**.** *Let $n \geq 4$. There exists a 3-restricted min-wise independent family $\mathscr{F} \subseteq S_n$ such that $|\mathscr{F}| \leq 3\sqrt{e}(1+o_n(1)) \cdot n(\log_2 n)^2$.* □

4.4 PSCAs and rankwise independent permutations

In this section we remark some recent results (of 2020 by [Yus20]) regarding the asymptotics of PSCAs and their isomorphic representation by rankwise independent families. We also state an open question posed in [Yus20] which we can later answer by using a respective result of [ITT00] for rankwise independent families. Afterwards we show how to improve upper bounds by comparison and combination of asymptotic/constructive results on PSCAs and rankwise independent families.

One can be interested in finding a PSCA(n,k,λ) having minimum multiplicity parameter λ – this automatically means that one is interested in minimizing the number of rows which is given by $\lambda k!$ (as previously discussed for SCAs in form of completely scrambling permutations).

Definition 4.4.1 (PSCA number)**.** *Let $n,k \in \mathbb{N}^\times$ with $n \geq k$. The smallest multiplicity $\lambda \in \mathbb{N}^\times$ such that there exists a PSCA over the alphabet $[n]$ of strength k and with multiplicity λ is denoted as $g(n,k)$.*

The quantity g is well-defined because the trivial choice of the entire symmetric group S_n as family will cover each k-subpermutation equally often (the resulting multiplicity of coverage is determined by Lemma 2.3.10).

For arbitrary strength k, Yuster was recently able to show the following lower bound.

Theorem 4.4.2 ([Yus20]). *If $k/2$ is a prime, then for all $n \geq k$ we have*

$$g(n,k) \geq \frac{\binom{n}{k/2} - \binom{n}{k/2-1}}{k!}.$$

For arbitrary k, provided n is sufficiently large in comparison to k, the bound $g(n,k) > n^{k/2-o_k(1)}$ applies. □

For the special case of strength $k = 3$ an upper bound (subquadratic in n) has been found. Moreover, a simple linear lower bound could be established.

Theorem 4.4.3 ([Yus20]). *Let $n \geq 3$. There exists a constant $C > 0$ such that*

$$\frac{n}{6} \leq g(n,3) \leq Cn(\log_2 n)^{\log_2 7}. \tag{4.56}$$

□

Problem 4.4.4. *In [Yus20], the following question is raised: "[...] $g(n,k)$ is lower bounded by a polynomial in n whose exponent grows with k. While it is not difficult to slightly improve upon the trivial upper bound $g(n,k) \leq n!/k!$, it would be interesting to obtain polynomial upper bounds for $g(n,k)$."*

We will indeed answer the question positively simply by pointing to a construction of polynomial size for rankwise independent permutations. That will give a better insight in the asymptotics of PSCAs, and in particular, will lead to an improved upper bound for $g(n,3)$ (following the asymptotic $C(1+o_n(1)) \cdot n \log_2 (n)^2$) in (4.56). In turn, classical results from design theory concerning PSCAs will be beneficial for the asymptotics of rankwise independent families.

We notice that the isomorphism in Lemma 2.3.13 allows to transfer bounds for the minimum cardinality of rankwise independent families to bounds for the minimum cardinalities of PSCAs (and hence we obtain bounds for the value $g(n,k)$ automatically).[5] We will state all subsequent results jointly with its implications for the bound of $g(n,k)$.

First, we point out the currently best known lower bound for PSCAs/rankwise independent families.

[5]Recall (see Lemma 2.3.13), that a d-family $\mathscr{A} \in \mathrm{PSCA}(n,k)$ has multiplicity $\lambda = d/k!$. Hence, renormalizing lower/upper bounds by a factor of $k!$ yields lower/upper bounds for $g(n,k)$.

Theorem 4.4.5 ([Bar04]). *Let $\mathscr{F} \subseteq S_n$ be a k-rankwise independent family. Then, with $!(\cdot)$ denoting the subfactorial (recall Section 2.1) we have*

$$g(n,k)k! \geq |\mathscr{F}| \geq \begin{cases} \sum_{i=0}^{h} !i\binom{n}{i}, & \text{if } k = 2h \\ \sum_{i=0}^{h} !i\binom{n}{i} + !(h+1)\binom{n-1}{h}, & \text{if } k = 2h+1 \end{cases}. \tag{4.57}$$

Consequently, $|\mathscr{F}| \geq \frac{!h}{h!} n^h (1 + o_n(1))$ if $k = 2h$, otherwise, if $k = 2h+1$, $|\mathscr{F}| \geq \frac{!h + !(h+1)}{h!} n^h (1 + o_n(1))$. □

Theorem 4.4.6 ([ITT00], cf. also [Iur22]). *Let $p \geq n \geq k$ such that p is a prime and $n \geq (k-1)!$. Furthermore, let $p_1 < \ldots < p_m$ be the sequence of all primes not exceeding $k-1$, i.e., $p_m \leq k-1$. Let $(e_1, \ldots, e_m) \in (\mathbb{N} \setminus \{0\})^m$ be a minimizer of $Q = \prod_{i=1}^{m} p_i^{e_i}$ under the side constraints that $(k-1)!$ divides Q, and $p_i^{e_i} > p$, for $i = 1, \ldots, m$. Then, there exists a k-rankwise independent family $\mathscr{P}$ of permutations of $[p]$ such that $|\mathscr{P}| \leq (p^k - p)Q^{\lfloor k/2 \rfloor}$. Consequently,[6] there exists a family $\mathscr{P}' \subseteq S_n$ satisfying*

$$g(n,k)k! \leq |\mathscr{P}'| = |\mathscr{P}| = n^{O(k^2/\ln k)}. \tag{4.58}$$

Remark 4.4.7. *The proof of Theorem 4.4.6 is constructive (the authors call it a "tie breaking scheme"). In case that $(k-1)! > n$, it is established that a k-rankwise independent family with cardinality of order $e^{O(k^3)}$ exists [ITT00].*

Example 4.4.8. *Let $k = 4$ and $p = n = 19$. The primes not exceeding 6 are $2, 3, 5$. We can choose, minimally, $(e_1, e_2, e_3) = (5, 3, 2)$, which ensures $6|2^5 3^3 5^2 = 21\,600$ and $\min\{2^5, 3^3, 5^2\} > 19$. Then, Theorem 4.4.6 implies existence of a 4-rankwise independent family $\mathscr{P}$ with $|\mathscr{P}| = (19^4 - 19) \cdot 21\,600^2 = 60\,793\,701\,120\,000$. We will refer in the following to a construction that is capable of generating much smaller families for the specific case of strength 4 (cf. Example 4.4.12 which shows the existence of a family of considerably smaller cardinality $24 \cdot 3\,139\,584$ for $n = 19$).*

As in [TIT03] consider for $L_q = (q_\ell, \ldots, q_1)$ (q is a fixed value), the sequence $C(\cdot)$, whose members are indexed by $q_1 + 1, q_2 + 1, \ldots, q_\ell + 1$, each index attaining a cardinality in $\mathbb{N}^\times$ and given recursively by

$$C(q_\ell + 1) = 8(q_{\ell-1}^2 + q_{\ell-1} + 1) q_{\ell-1} C(q_{\ell-1} + 1). \tag{4.59}$$

By considering special projective plains in finite geometry, the following result is obtained.

Lemma 4.4.9 ([TIT03, Theorem 5]). *Let $L_q = (q_\ell, \ldots, q_1)$ and $C(\cdot)$ be as before. Then, for $i \in \{2, \ldots, \ell\}$, the existence of a 4-rankwise independent $C(q_{i-1}+1)$-family of permutations of $[q_{i-1}+1]$ implies the existence of a 4-rankwise independent $C(q_i+1)$-family of permutations of $[q_i + 1]$.* □

[6] Cf. Proposition 2.2.7.

Remark 4.4.10 (cf. also [Iur22]). *The following optimization is similar to Remark 4.3.11. Tarui et al. [TIT03] make use of the previous Lemma 4.4.9 to build* 4*-rankwise independent families employing as base case the* 120*-family* S_5 *(trivially satisfying the requirement), and afterwards to obtain by estimation an upper bound. We observe, however, that by Levenshtein's construction we can come up with a* $\mathrm{PSCA}(5,4,1)$ *and map it via Lemma 2.3.13 to a* 4*-rankwise independent family of just* 24 *permutations. As the recursion for* $C(\cdot)$ *is homogeneous in the cardinality of the family employed as base case, we automatically obtain a bound sharpened by a factor of* 5*.*

Theorem 4.4.11 (Estimate of [TIT03] sharpened by factor 5). *Let* $n \geq 4$*. Then, there exists a 4-rankwise independent family* $\mathscr{F} \subseteq S_n$*, with cardinality*

$$g(n,4) \cdot 4! \leq |\mathscr{F}| \leq 3e(1+o_n(1))n^3(\log_2 n)^6. \tag{4.60}$$

□

For transparency, in the following Example 4.4.12, we enclose the straightforward derivation of $g(n,4)$-values for small range of n; successively the same procedure is repeated for the $g(n,3)$-values obtainable from the recursion in Remark 4.3.11 of the previous section.

Example 4.4.12 (Small scale evaluation of recursion for 4-rankwise independent families). *By Remark 4.4.10, we have* $g(5,4) = 1$*. Consider* $L_{16} = (q_3, q_2, q_1) = (16, 8, 4)$*, i.e.,* $\ell = 3$*. Then, using* (4.59)*,* $C(q_2+1) = C(9) = 8(q_1^2+q_1+1)q_1C(q_1+1) = 672 \cdot 24$*. Lastly, we obtain* $C(q_3+1) = C(17) = 8(q_2^2+q_2+1)q_2C(q_2+1) = 3139584 \cdot 24$*. Alternatively, consider* L_{64} *such that, as before* $C(9) = 672 \cdot 24$*, and consequently* $C(65) = 8 \cdot (8^2+8+1)(8)C(9) = 3\,139\,584 \cdot 24$*. Consider also as last alternative the sequence* L_{256}*, for which* $C(17) = 672 \cdot C(5) = 672 \cdot 24$*, and consequently* $C(257) = 23\,482\,368 \cdot 24$*.*

Example 4.4.13 (Small scale evaluation of recursion for 3-restricted min-wise independent families). *We can rely on the base case* $C(4) = 6$ *by Remark 4.3.11. For* $L_{16} = (q_3, q_2, q_1) = (16, 8, 4)$*, applying the recursion for the cardinalities pointed out in Remark 4.3.11 implies* $C(q_2) = C(8) = 2(q_1+1)C(q_1) = 10 \cdot 6$*, and consequently* $C(16) = 2 \cdot (8+1) \cdot 10 \cdot 6 = 180 \cdot 6$*. Employing the alternative sequence* $L_{64} = (64, 8, 4)$*, we get* $C(64) = 2 \cdot (8+1) \cdot 60 = 180 \cdot 6$*. Finally, when choosing* L_{256} *we obtain* $C(16) = 10 \cdot 6$ *and therefore* $C(256) = 340 \cdot 6$*.*

In the previous examples, for each number n, there is an optimal choice for L_q leading to the smallest family obtainable via the recursive constructions (hereby we use the monotonicity of PSCAs, see Lemma 2.3.10, to obtain bounds for smaller and more general n). Smallest herewith obtainable upper bounds for the resulting $g(n,4)$-values are assembled in Table 4.1. To get an impression of the quality of these bounds, these are stated jointly with other values established in the literature. In Table 4.1 we can observe that the considerations in Remark 4.4.10 lead to an improvement of the number $g(n,4)$ obtained in [GW22] for five values. Hereby, the bounds are reduced by a factor of up to 7.5. Afterwards we assemble in Table 4.2 the values obtained for bounds of $g(n,k)$.

n	Bound in [GW22] (best known)	Bound by optimized recursive construction in Remark 4.4.10
5	1	1
6	1	672
7	2	672
8–12	18	–
13	234	–
14–17	5040	**672**
18–21	5040	3 139 584
22	18 480	–
23	425 040	–
24	10 200 960	**3 139 584**
25-65	/	3 139 584
66-257	/	23 482 368

Table 4.1: Upper bounds for $g(n, 4)$ (cf. [Iur22]). Entries beating best-known values are in bold print.

n	Best known bound (with reference)		Bound by optimized recursive construction in Remark 4.3.11
4		1	1
5–7	[Yus20];[NJL22, GW22]	2	10
8	[GW22]	3	_
9	[NJL22]	4	_
10–12	[GW22]	6	_
13–14	[GW22]	7	_
15–16	[GW22]	16	**10**
17–19	[GW22]	19	180
20–32	[GW22]	96	180
33–64		/	180
65–256		/	340

Table 4.2: Upper bounds for $g(n, 3)$. Entries beating best-known values are in bold print.

Remark 4.4.14. *Lastly, for completing the discussion on asymptotics, we anticipate some considerations on the Levenshtein construction in Chapter 6: We remark that* $g(n, n-1) = 1$*; in other words, the super-exponential amount of* $(n-1)!$ *permutations is enough to host an optimal PSCA of strength* $n-1$ *(of multiplicity* 1*).*

4.5 Comparison of bounds

We collect in Table 4.3 best bounds encountered in the present chapter for differently strong "forms of k-scramblingness" that families of permutations can satisfy. It should be noticed that the bounds are asymptotically sharp, i.e., $\Theta(\log_2 \log_2 n)$ (respectively $\Theta(\log_2 n)$) for (completely) k-scrambling permutations, whereas for sizes of PSCAs/min-wise independent families with fixed parameter k there is an asymptotic gap of polynomial magnitude.

Property	Lower bound	Upper bound
k-scrambling	$(1+o_n(1))\log_2 \log_2 n$	$k2^k(1+o_n(1))\log_2 \log_2 n$
3-scrambling	$(1+o_n(1))\log_2 \log_2 n$	$(1+o_n(1))\log_2 \log_2 n$
completely k-scrambling	$\frac{(k-1)!}{\log_2 e}(1+o_n(1))\log_2 n$	$\frac{k}{\log_2 \frac{k!}{k!-1}}(1+o_n(1))\log_2 n$
completely 3-scrambling	$2\ln(2)(1+o_n(1))\log_2 n$	$2(1+o_n(1))\log_2 n$
k-restr. min-wise independ.	$\frac{1+o_n(1)}{\lfloor k/2 \rfloor^{\lfloor k/2 \rfloor}} n^{\lfloor k/2 \rfloor}$	$2^k((k-1)!)^k\, n^k$
3-restr. min-wise independ.	———	$\mathbf{3}\sqrt{e}(1+o_n(1))\cdot n(\log_2 n)^2$
4-restr. min-wise independ.	———	$\mathbf{6}\sqrt{e}(1+o_n(1))\cdot n(\log_2 n)^3$
min-wise independ.	$e^{n-o_n(1)}$	$e^{n-o_n(1)}$
strength-k PSCA (size)	$\frac{!\lfloor k/2 \rfloor}{\lfloor k/2 \rfloor!}(1+o_n(1))n^{\lfloor k/2 \rfloor}$	$n^{O(k^2/\ln k)}$
strength-3 PSCA (size)	———	$\mathbf{3}\sqrt{e}(1+o_n(1))\cdot n(\log_2 n)^2$
strength-4 PSCA (size)	———	$\mathbf{3}e(1+o_n(1))n^3(\log_2 n)^6$
strength-$(n-1)$ PSCA (size)	$(n-1)!$	$(n-1)!$

Table 4.3: Summary of bounds encountered in this chapter. Dashes indicate that the more general bound (arbitrary k) should be consulted. Coefficients in bold print highlight improvements pointed out. The results stated for PSCAs are valid for rankwise independent families, too.

CHAPTER 5

Constructions of completely scrambling families

In the present chapter, we focus on construction algorithms aiming to generate completely scrambling families of permutations small cardinality. It is convenient to describe some of the constructions in the language of SCAs. The development of constructions for the regularized case is discussed in the language of PSCAs in the next chapter. A dedicated section discusses a particularly convenient and elegant construction technique due to Tarui [Tar08] for the special case of strength three. The chapter is concluded with a brief summary of other approaches proposed throughout the literature.

5.1 Deterministic polynomial time construction keeping asymptotic bounds

This section deals with a construction method due to Chee et al. [CCHZ13] which provides a deterministic polynomial time algorithm for generating actual instances of completely k-scrambling families (hereby Spencer's asymptotic upper bound (4.14) will not be exceeded). In fact, the upper bound derived via the probabilistic method in Chapter 4 does not provide information on how to come up with completely k-scrambling families. We make explicit some argumentation steps which are kept short in [CCHZ13].

Some auxiliary results will be first explained. We make use of the set

$$\mathrm{COV}_q := \{\pi \in S_n : q \text{ is contained as subsequence in } \pi\}\,. \tag{5.1}$$

Lemma 5.1.1. *Consider $A \subseteq S_{n,k}$. The expected number of members of A, which are contained as subsequence by a permutation $\pi \in S_n$ chosen uniformly at random, equals $|A|/k!$.*

Proof. Formally, this can be seen by

$$\mathop{\mathbb{E}}_{\pi\in S_n}\left[\sum_{q\in A}\mathbb{1}_{\mathrm{COV}_q}(\pi)\right]=\sum_{q\in A}\mathop{\mathbb{E}}_{\pi\in S_n}\left[\mathbb{1}_{\mathrm{COV}_q}(\pi)\right]=\sum_{q\in A}\frac{1}{k!}=\frac{|A|}{k!}. \tag{5.2}$$

□

Let $A\subseteq S_{n,k}$ be fixed. For $m\leq n$ and $r\in S_{n,m}$, consider the quantity

$$\mathrm{EC}(r):=\mathop{\mathbb{E}}_{\pi\in S_n\cap\mathrm{COV}_r}\left[\sum_{u\in A}\mathbb{1}_{\mathrm{COV}_u}(\pi)\right], \tag{5.3}$$

which captures the expected number of sequences in A that are contained as subsequence by a permutation randomly drawn from COV_r (with uniform probability).

Lemma 5.1.2. *Let $u=(u_1,\ldots,u_k)\in S_{n,k}$, $r=(r_1,\ldots,r_m)\in S_{n,m}$. Let $\alpha_1,\ldots,\alpha_\ell$ be pairwise distinct symbols satisfying firstly,*

$$\{\alpha_1,\ldots,\alpha_\ell\}=\{u_1,\ldots,u_k\}\cap\{r_1,\ldots,r_m\} \tag{5.4}$$

and secondly, that $(\alpha_1,\ldots,\alpha_\ell)$ is a subsequence of u. Then,

$$\mathop{\mathbb{E}}_{\pi\in S_n\cap\mathrm{COV}_r}\left[\mathbb{1}_{\mathrm{COV}_u}(\pi)\right]=\begin{cases}\frac{\ell!}{k!}, & \text{if } (\alpha_1,\ldots,\alpha_\ell) \text{ is subsequence of } r\\ 0, & \text{otherwise}\end{cases}. \tag{5.5}$$

Proof. If u and r place the α_i in a different order, then clearly $\mathrm{COV}_u\cap\mathrm{COV}_r=\emptyset$ and the expectation vanishes. Otherwise, the expectation is taken over permutations containing r as subsequence, and consequently containing $(\alpha_1,\ldots,\alpha_\ell)$ as subsequence. In order to contribute to the expectation, the remaining symbols of u have to occur between suitable α_j's and in the correct order among themselves; this occurs with probability $\frac{1}{(\ell+1)(\ell+2)\cdots(\ell+k-\ell)}=\ell!/k!$. □

Theorem 5.1.3. *Let $A\subseteq S_{n,k}$. There exists a sequence $Q_1,\ldots,Q_n$ whose members lie in $\bigcup_{i\geq 1}S_{n,i}$ and satisfy the following requirements.*

(i) $Q_i\in S_{i,i}\subseteq S_{n,i}$, $i=1,\ldots,n$.

(ii) Q_i *is a subsequence of* Q_{i+1}, $i=1,\ldots,n-1$.

(iii) $\mathrm{EC}(Q_i)\leq\mathrm{EC}(Q_{i+1})$, $i=1,\ldots,n-1$.

Proof. Q_1 is forced to be the singleton tuple (1). Suppose we have $Q_i\in S_{i,i}$. We wish to find $Q_{i+1}\in S_{i+1,i+1}$ satisfying (iii). Therefore, it remains to figure out which are feasible index positions j, $j=0,\ldots,n$, after which the symbol $i+1$ can be inserted in Q_i in order

to obtain a candidate for Q_{i+1}. Denote by $\{C_0, \ldots, C_i\} \subseteq S_{i+1,i+1}$ the corresponding set of candidates. We show that any choice

$$Q_{i+1} \in \operatorname{argmax}_{C \in \{C_0,\ldots,C_i\}} \mathrm{EC}(C) \tag{5.6}$$

meets the requirement (iii). In fact,

$$\begin{aligned}
\mathrm{EC}(Q_i) &= \mathop{\mathbb{E}}_{\pi \in S_n \cap \mathrm{COV}_{Q_i}} \left[\sum_{u \in A} \mathbb{1}_{\mathrm{COV}_u}(\pi) \right] \\
&= \frac{i!}{n!} \left(\sum_{(\pi,u) \in \mathrm{COV}_{C_0} \times A} \mathbb{1}_{\mathrm{COV}_u}(\pi) + \ldots + \sum_{(\pi,u) \in \mathrm{COV}_{C_i} \times A} \mathbb{1}_{\mathrm{COV}_u}(\pi) \right) \\
&= \frac{i!}{(i+1)!} \left(\frac{(i+1)!}{n!} \sum_{(\pi,U) \in \mathrm{COV}_{C_0} \times A} \mathbb{1}_{\mathrm{COV}_u}(\pi) + \ldots \right. \\
&\qquad \left. + \frac{(i+1)!}{n!} \sum_{(\pi,u) \in \mathrm{COV}_{C_i} \times A} \mathbb{1}_{\mathrm{COV}_u}(\pi) \right) \\
&= \frac{1}{i+1} \sum_{j=0}^{i} \mathrm{EC}(C_j).
\end{aligned}$$

This means that there must exist an index (especially a maximizing index) j such that $\mathrm{EC}(Q_i) \leq \mathrm{EC}(C_j)$. □

Corollary 5.1.4. *For $A \subseteq S_{n,k}$ it is possible to find in polynomial time (polynomial in n) a permutation $\pi \in S_n$ covering a fraction of at least $\frac{1}{k!}$ members of A.*

Proof. Executing the steps in the proof of Theorem 5.1.3 leads to the element $Q_n \in S_{n,n}$ which by transitivity satisfies $\mathrm{EC}(Q_n) \geq \mathrm{EC}(Q_1) \geq \frac{1}{k!}|A|$ (the lowest bound is a consequence of evaluating (5.3) for $r = (1) \in S_{n,1}$, cf. also (5.2)). It remains to show how the calculation of the values $\mathrm{EC}(C_j)$ can be done efficiently, in order to find a maximizer (5.6) in polynomial time. This can, however, be achieved by noticing that

$$\mathrm{EC}(C_j) = \mathop{\mathbb{E}}_{\pi \in S_n \cap \mathrm{COV}_{C_j}} \left[\sum_{u \in A} \mathbb{1}_{\mathrm{COV}_u}(\pi) \right] = \sum_{u \in A} \mathop{\mathbb{E}}_{\pi \in S_n \cap \mathrm{COV}_{C_j}} [\mathbb{1}_{\mathrm{COV}_U}(\pi)] \tag{5.7}$$

is a sum of $|A| \leq \frac{n!}{(n-k)!} \leq n^k$ summands each of which can be determined in linear time by employing the identity (5.5) per summand. Such a sum has to be calculated at most for n candidate permutations C_j.

□

At this point, all preparatory steps for an efficient construction have been concluded. We now put them together.

Theorem 5.1.5 ([CCHZ13]). *Let $n \geq k$, where k is fixed. There exists a polynomial time (polynomial in n) algorithm to construct a d-family in* $\mathrm{SCA}(n,k)$[1] *such that Spencer's upper bound is maintained, i.e., $d \leq k/\log_2 \frac{k!}{k!-1} \cdot \log_2 n + 1$.*

Proof. The procedure is given in Algorithm 1. It remains just to check its run time: The auxiliary routine FindBestCandidate runs in polynomial time by Corollary 5.1.4. The size of the output family will depend on the number of iterations in the while loop of the algorithm. We show that the latter number is only logarithmic in n: After the ℓ-th iteration, by Corollary 5.1.4, it will be granted $|A| \leq \left(\frac{k!-1}{k!}\right)^{\ell} |S_{n,k}|$, which is strictly smaller than 1 certainly as soon as $\ell > \log_2 \left(\frac{n!}{(n-k)!}\right) / \log_2 \left(\frac{k!}{k!-1}\right)$. The returned family will possess cardinality $d \leq \lceil \log_2 (n!/(n-k)!) / \log_2 (k!/(k!-1)) \rceil$. Therefore, Spencer's bound is maintainable deterministically within polynomial overall run time. □

Remark 5.1.6. *We emphasize that in the previous proof the incorporation of the fact that $|A|$ is integral by iteration leads to slightly sharpened bounds (cf. [CCHZ13]). As experiments in [CCHZ13] show, typically, maximizers slightly exceeding the expectations are found within the iterations and therefore, in praxis, even lower bounds result from the entire constructive procedure.*

Another observation due to [CCHZ13] is that "iterations with reversals" asymptotically lead to a slight improvement. Hereby, one would replace line 6 of Algorithm 1 by $\mathscr{F} \leftarrow \mathscr{F} \cup \{\pi, \pi \circ \rho\}$ (with the reversing permutation $\rho(i) = n-i+1$). The successive line in the algorithm would be adapted accordingly in order to add to D also the k-sequences covered by $\pi \circ \rho$. The idea goes back to [KHL+12] and exploits the fact that reversion allows to obtain per iteration a second maximizer without additional computational costs. Surprisingly, an empirical study due to [CCHZ13] shows that for actual construction purposes, when $k = 3$, $n \leq 90$, this supposed improvement actually produces families of larger sizes. For $k \in \{4, 5\}$, $n \leq 90$, on the other hand, a profit in size reduction seems almost consistently to emerge – the gain, however, seems of negligible magnitude. It is still unsolved what underlying mechanisms this paradoxical behavior for small n is due to (cf. [CCHZ13]). The adaptation of Theorem 5.1.5 incorporating reversals leads to the following bound.

Theorem 5.1.7 ([CCHZ13]). *Let $n \geq k$, where k is fixed. There is an algorithm, whose runtime is polynomial in n, to generate a d-family of completely k-scrambling permutations with*

$$N^*(n,k) \leq d \leq 2 \log_2 \left(\frac{n!}{(n-k)!}\right) / \log_2 \left(\frac{k!}{k!-2}\right).$$

□

[1] By Lemma 2.2.9 we can easily (linear time) convert it to a completely k-scrambling family.

Algorithm 1 Deterministically keeping Spencer's bound

Input: Permutation length $n \in \mathbb{N}^{\times}$, strength $k \leq n$
Output: SCA of strength k over the symbol set $[n]$

procedure DERANDOMIZEDSPENCER(n, k)
    $\mathscr{F} \leftarrow \emptyset$
    $A \leftarrow S_{n,k}$
    while $A \neq \emptyset$ **do**
        $\pi \leftarrow$ FINDBESTCANDIDATE(A)
        $\mathscr{F} \leftarrow \mathscr{F} \cup \{\pi\}$
        $D \leftarrow \{x \in S_{n,k} : x \text{ is contained as subsequence in } \pi\}$
        $A \leftarrow A \setminus D$
    return $\mathscr{F}$

procedure FINDBESTCANDIDATE(A)
    $w = (1)$ ▷ (tuple of length 1)
    for $i = 2, \ldots, n$ **do**
        for $j = 0, 1, \ldots, i$ **do**
            Let C_j, be the word resulting from w by inserting symbol i in w directly after position j.
            Evaluate $\mathrm{EC}(C_j)$, according to (5.7) using simplification (5.5)
        $w \leftarrow \mathrm{argmax}_{C \in \{C_0, \ldots, C_i\}} \mathrm{EC}(C)$ ▷ Pick first (or an arbitrary) maximizer
    return w

5.2 Tarui's construction for strength three

Despite the fact that (completely) k-scrambling permutations and their properties were already analyzed in the 1950s (cf. [Dus50]) respectively 1970s (cf. [Spe71]), it is rather surprising that the following construction technique for $k = 3$ due to J. Tarui was only discovered in 2005. For higher values of k no comparable constructive approaches seem to be available and it is an open question to find such ones.

Theorem 5.2.1 (Tarui, [Tar08])**.** *For fixed $q \in \mathbb{N}^{\times}$, let $f(q)$ denote the maximum value attainable by n such that there exists a completely 3-scrambling q-family of permutations of $[n]$. Then, $f(q) \geq \binom{\lfloor q/2 \rfloor}{\lfloor q/4 \rfloor}$.*

Proof. Let $r := \lfloor q/2 \rfloor$ and choose $U = \{A_1, \ldots, A_m\} \in 2^{\{1,\ldots,r\}}$ to be an m-set satisfying the antichain property with respect to set inclusion. By choosing $U := \binom{[r]}{\lfloor r/2 \rfloor}$ we assure ourselves that U can be chosen at least of size $\binom{r}{\lfloor r/2 \rfloor}$. In the following we introduce $2r$ strict linear orders on the set U: The first group of strict linear orders, $<_x$, for

$x = 1, \ldots, r$, is determined by

$$A_i <_x A_j :\Leftrightarrow \begin{cases} & (x \in A_i \wedge x \notin A_j) \\ \vee & (x \in A_i \cap A_j \wedge i < j) \\ \vee & (x \notin A_i \cup A_j \wedge i < j) \end{cases}, \tag{5.8}$$

whereas the second group, $\prec_x$, for $x = 1, \ldots, r$, is given by

$$A_i \prec_x A_j :\Leftrightarrow \begin{cases} & (x \in A_i \wedge x \notin A_j) \\ \vee & (x \in A_i \cap A_j \wedge i > j) \\ \vee & (x \notin A_i \cup A_j \wedge i > j) \end{cases}. \tag{5.9}$$

The only difference within the defining logical disjunctions for $<_x$ and $\prec_x$ is the orientation of the "$<$" sign appearing in the second and third operand. By their nature, $<_x$ and $\prec_x$ are strict linear orders.

The claim is now that $<_x$ and $\prec_x$, read as permutations of U, collectively form a completely 3-scrambling $(2r)$-family of permutations of $[m]$. This is equivalent to showing that there is a linear ordering ρ among $<_1, <_2, \ldots, <_r, \prec_1, \prec_2, \ldots, \prec_r$ such that $A_i \; \rho \; A_j \; \rho \; A_k$ applies for an arbitrary index selection $(i, j, k) \in S_{m,3}$. This fulfillment of the latter claim is shown as follows: By the antichain property, the existence of an element $x \in A_i \setminus A_k$ is granted. We claim that either according to $<_x$, or according to $\prec_x$, the sequence (A_i, A_j, A_k) forms an ascending chain. Indeed, if $x \in A_j$, we have that (A_j, A_k) is ascending for both orders. However, to obtain that also (A_i, A_j) is ascending, we have to pick $<_x$ iff $i < j$.

The remaining case $x \notin A_j$ is handled accordingly and the assertion ensues. □

Remark 5.2.2. *We observe that Tarui's construction cannot be optimized by individuation of a longer antichain in the previous proof: Indeed, Sperner's theorem (see [And87, Theorem 1.2.1]) affirms that any antichain of* $[n]$ *(with respect to set inclusion) of maximum cardinality corresponds to* $\binom{[n]}{\lfloor n/2 \rfloor}$ *or* $\binom{[n]}{\lceil n/2 \rceil}$.

From Theorem 5.2.1, compare [Tar08], we get the improved logarithmic upper bound for completely 3-scrambling families of permutations already mentioned in Theorem 4.2.3. Moreover, Tarui was able to prove the following fact (where, however, the value of the limit is still unknown).

Theorem 5.2.3 ([Tar08])**.** *The limit* $\lim_{n \to \infty} \frac{N^*(n,3)}{\log_2(n)}$ *exists.* □

In Algorithm 2 we provide a procedure useful to actually generate scrambling permutations following the strategy presented in the proof of Theorem 5.2.1.

Algorithm 2 Generating 3-scrambling permutations via Tarui's construction

Input: Permutation length $n \in \mathbb{N}^{\times}$
Output: Completely 3-scrambling family of permutations of $[n]$

procedure TARUI(n)
  $q \leftarrow \min\left\{q \in \mathbb{N} : \binom{\lfloor q/2 \rfloor}{\lfloor q/4 \rfloor} \geq n\right\}$ ▷ Determine the size of the output family
  $r \leftarrow \lfloor q/2 \rfloor$
  $S \leftarrow \{c \in [q] : c > n\}$ ▷ Keep track of superfluous symbols

  ▷ Construct the "universe", a set on which we want to consider permutations

  $U \leftarrow \{A \subseteq [r] : |A| = \lfloor r/2 \rfloor\} \subseteq 2^{[r]}$ ▷ Technically, consider U as ordered list $[A_1, \ldots, A_\ell]$ with $\ell = \binom{r}{\lfloor r/2 \rfloor}$, where the listed items are lexicographically ordered

  $L \leftarrow \{\}$ ▷ Initialize empty list of permutations
  for $x = 1, \ldots, r$ **do**
    $B_x \leftarrow [i : x \in A_i, i = 1, \ldots, \ell]$ ▷ (ordered list)
    $C_x \leftarrow [i : x \notin A_i, i = 1, \ldots, \ell]$ ▷ (ordered list)

    ▷(Standard concatenation of lists and the reversal of a list is used below)
    $D_x \leftarrow \text{CONCAT}(B_x, C_x)$
    $D'_x \leftarrow \text{CONCAT}(\text{REV}(B_x), \text{REV}(C_x))$
    $L \leftarrow L \cup \{D_x, D'_x\}$
  Eliminate from each of the elements of L the occurrences of $c \in S$

  return $\{\pi^{-1} : \pi \in L\}$ ▷ If the output is preferred as SCA, then inversion has to be omitted here.

5.3 Generation of k-scrambling permutations: further approaches

In this section, we collect approaches having already been pursued in the literature to provide a way to generate completely k-scrambling permutations. As the interest has shifted from completely scrambling permutations to sequence covering arrays due to their practical applicability, most recent construction techniques are tailored to the latter concept.

In [BTI12, BEI+12] permutations are handled as linear orders and processed via their incidence matrices. On these matrices of Boolean entries, Boolean constraints determining a strict linear order are imposed (irreflexibility, asymmetry, and transitivity). When comparing any two symbols $x, y \in [n]$, the order relation has hereby the function to determine which is the lefter-most element in the notation of the represented permutation as tuple.

Within this framework a fixed amount of initially kept indeterminate permutations, i.e., incidence structures can be provided as initial guess. Answer set programs additionally incorporating constraints enforcing (here in the language of SCAs) the coverage of each $s \in S_{n,k}$ by at least one of the incidence structures are designed and passed to suitable solving software. When no solution is found during the solving process, the amount of initially guessed incidence structures is successively increased. This approach is mixed with a greedy strategy in [BEI$^+$12].

A couple of bio-inspired heuristic algorithms have been applied (heuristics here address the problem how to promisingly choose the next permutation within a "one permutation at a time" procedure as Algorithm 1): A search via a "bees algorithm" in [MKR12], a "fish swarm algorithm" is presented in [RSK$^+$20], and a "elitist-flower pollination"-based strategy in [NZAA18]. Apart from these works it seems that methods dealing with strength five (or higher strengths) are rarely represented in literature.

An interesting finding is a product construction due to [CCHZ13]: For strength $k = 3$, if one has a so-called *properly signed* d-family of completely 3-scrambling permutations of $[n]$ and a second properly signed $\tilde{d}$-family of completely 3-scrambling permutations of $[\tilde{n}]$, then a $(d + \tilde{d})$-family of completely 3-scrambling permutations of $[n\tilde{n}]$ can be constructed from them (being properly signed as well). Starting from small exemplars, this approach can be used to generate larger instances outperforming in terms of optimality of cardinality all methods described in the current section [CCHZ13]. Even Tarui's construction being highly efficient could be beaten as soon as $n \geq 40$.

Remark 5.3.1. *It seems to be hard to find* optimal *SCAs, i.e., on N^* rows. In [BEI$^+$12] a variant of the problem formulation is considered (called "generalized event sequence testing problem") and shown to be NP-complete. In [CCHZ13], variations of the problem requiring the satisfaction of additional side constraints, such as prohibiting coverage of certain pattern of subsequences are classified as difficult to solve by pointing e.g. to the NP-complete* betweenness problem *(cf. [Yan82]).*

CHAPTER 6

Constructions of Perfect Sequence Covering Arrays

6.1 Construction via Varshamov-Tenengolts codes

Investigating the consequences of his study of the Varshamov-Tenengolts codes in [Lev91], Levenshtein noted connections to existence assertions for objects in design theory (such as directed and undirected Steiner systems). We rely here on the presentation in [Na21], which contains Levenshtein's conclusions restricted and tailored to the setting of PSCAs.

As special case of so-called *double order relations* appearing in [Lev91], in [Na21] the following notion is introduced.

Definition 6.1.1. *Let $\sigma = (\sigma_1, \ldots, \sigma_n) \in S_n$. Let us write $F(\sigma) = (f_1, \ldots, f_{n-1}) \in B_2^{n-1}$ for the $(n-1)$-tuple of Boolean values f_j, answering if (σ_j, σ_{j+1}) is a falling sequence, i.e.,*

$$f_j = \begin{cases} 1, & s_j > \sigma_{j+1} \\ 0, & s_j < \sigma_{j+1} \end{cases}.$$

We note that, in general, $F(\sigma)$ does not reveal σ, as simple counterexamples can be found. However, the following elementary properties hold.

Lemma 6.1.2. *Consider $\sigma, \tau \in S_n$ and $r \in S_{n,n-1}$ such that r is a subsequence of σ. The following assertions are valid.*

(i) $F(\sigma) \neq F(\tau)$ *implies* $\sigma \neq \tau$.

(ii) $F(r) \in \{0,1\}^{n-2}$ *is a subsequence of* $F(\sigma) \in \{0,1\}^{n-1}$.

Proof. (i) is clear. For (ii), if r arises from σ by deletion of one entry, say the one at position i, we have

$$\sigma = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_i, \sigma_{i+1}, \ldots, \sigma_n) \text{ and } r = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_n) \tag{6.1}$$

with corresponding F-values ($c \in \{0, 1\}$)

$$F(\sigma) = (f_1, \ldots, f_{i-1}, f_i, f_{i+1}, \ldots, f_{n-1}) \text{ and } F(r) = (f_1, \ldots, f_{i-2}, c, f_{i+1}, \ldots, f_{n-1}). \tag{6.2}$$

We make a short case distinction on the monotony of the sequence $(\sigma_{i-1}, \sigma_i, \sigma_{i+1})$:

In case $(\sigma_{i-1}, \sigma_i, \sigma_{i+1})$ is monotonic, by transitivity, it will as well be $(\sigma_{i-1}, \sigma_{i+1})$. Therefore, $c = f_{i-1} = f_i$ and $F(\sigma)$ covers $F(r)$.

For the remaining non-monotonic cases, $\{f_{i-1}, f_i\} = \{0, 1\}$ and therefore, regardless which value $c \in \{0, 1\}$ attains, again coverage of $F(r)$ by $F(\sigma)$ ensues. □

Theorem 6.1.3. *Consider the partition of the symmetric group S_n given by the parts*

$$U_a := \left\{\sigma \in S_n : F(\sigma) \in \mathrm{VT}^{n-1,a}\right\}, \quad a = 0, \ldots, n-1. \tag{6.3}$$

Then, for arbitrary a, the elements of U_a form an instance of $\mathrm{PSCA}(n, n-1, 1)$.

Proof. Let n and a be fixed. First, we observe that no element $s \in S_{n,n-1}$ is covered more than once: Seeking a contradiction, assume two distinct $y, z \in U_a \subseteq S_{n,n}$ both cover the tuple s. Then, by Lemma 6.1.2, $F(s)$ is a subsequence of $F(y)$ and of $F(z)$ (forming two different binary words). This is impossible, as $F(y), F(z) \in \mathrm{VT}^{n-1,a} \subseteq B_2^{n-1}$ and $\mathrm{VT}^{n-1,a}$ is a B_2^{n-2}-perfect code correcting single deletions (see Theorem 3.3.7).

Each $u \in U_a$ will cover n subsequences $s \in C_u \subseteq S_{n,n-1}$. Moreover, the sets of covered subsequences C_u, $C_{\tilde{u}}$ are mutually disjoint for differing $u, \tilde{u}$. Consequently, the members of U_a cover a proportion of $\frac{|U_a|n}{n!}$ of the subsequences in $S_{n,n-1}$. To show the claim it remains to prove $|U_a| = (n-1)!$.

The proof of the latter relies on the collective behavior of U_a, for all $a = 0, \ldots, n-1$. By contradiction, if there was $\tilde{a} \in \{0, \ldots, n-1\}$ such that $|U_{\tilde{a}}| < (n-1)!$, then the following conflict would arise:

$$\begin{aligned} n! = \sum_{a=0}^{n-1} |U_a| &= |U_{\tilde{a}}| + \sum_{a \neq \tilde{a}} |U_a| && (6.4) \\ &< (n-1)! + \sum_{a \neq \tilde{a}} |U_a| \\ &\leq (n-1)! + (n-1)(n-1)! = n!. && (6.5) \end{aligned}$$

Hereby, in (6.4) we used the partitioning property of $\mathrm{VT}^{n-1,a}$, and in (6.5) the fact that, by the injectivity shown in the first part of the proof, $|U_a| \leq (n-1)!$. □

Remark 6.1.4. *According to [MvT99], in the 1990s, Levenshtein conjectured (originally referring to Steiner systems, see Section 2.3) that a* $\mathrm{PSCA}(n,k,1)$ *exists iff* $n \in \{k, k+1\}$ *(*$k \geq 3$*). In [MvT99] the conjecture was computationally falsified by constructing a counterexample with* $n = k + 2 = 6$*.*

Example 6.1.5. *The condition* (6.3) *of the previous theorem provides a precise instruction on how to construct a PSCA. We can take the Varshamov-Tenengolts code* $\mathrm{VT}^{4,0}$ *from Example 3.3.6 which yields the subsequent* $\mathrm{PSCA}(5,4,1)$*. The set of all permutations from* S_5 *whose F-sequence* $(f_1,\ldots,f_4)$ *coincides with the first codeword* $0000 \in \mathrm{VT}^{4,0}$ *is given by* $\{(1,2,3,4,5)\}$ *and is reflected in the first row in the below matrix; the remaining three codewords are represented by the groups of rows with index scope* 2-12*,* 13-23*, respectively* 24-24*.*

$$\left[\begin{array}{c|cccccccccc|cccccccccc|c}
1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 & 5 & 5 & 5 & 5 \\
2 & 4 & 5 & 5 & 4 & 5 & 5 & 4 & 5 & 5 & 5 & 5 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 2 & 4 \\
3 & 3 & 3 & 4 & 3 & 3 & 4 & 2 & 2 & 4 & 2 & 3 & 3 & 4 & 2 & 4 & 4 & 2 & 3 & 3 & 2 & 3 & 3 & 3 \\
4 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 4 & 4 & 4 & 2 \\
5 & 5 & 4 & 3 & 5 & 4 & 3 & 5 & 4 & 2 & 3 & 2 & 4 & 3 & 4 & 2 & 1 & 3 & 2 & 1 & 3 & 2 & 1 & 1
\end{array}\right]^{\mathsf{T}}$$

6.2 Computational constructions for PSCAs

As the previous chapter discussed only PSCAs for the case $k = n-1$, we now turn to the more general case, for which –according to the current state of knowledge– optimal PSCAs can only be found by expensive computational backtracking approaches. In [MvT99] Mathon and van Trung, relying on a backtracking algorithm of [Mat97], computed various combinatorial designs including $\mathrm{PSCA}(n,k,\lambda)$ with $\lambda = 1$. In particular, two non-equivalent examples of a $\mathrm{PSCA}(6,4,1)$ are found by this approach – moreover, they figured out that for those parameters $(n,k,\lambda) = (6,4,1)$ only PSCAs equivalent to one the two found examples exist. The basic framework of their algorithm is quite simple: As for some of the seen approaches for the generation of scrambling permutations, it is tried to build the PSCA permutation by permutation. For this, whenever a permutation is added, it is checked, for each of the $\binom{n}{k}$ sequences $s \in S_{n,k}$ hereby getting covered, if the coverage is novel. If the answer is affirmative, the check is passed and, recursively, all candidates for the next rows are processed accordingly.

It is immediate how the algorithm can be generalized to search for $\mathrm{PSCA}(n,k,\lambda)$, $\lambda \geq 1$; this is presented in [Na21] with an extensive discussion and some details among which e.g. an appropriate way to keep track of the coverage of subsequences via incidence vectors.

The most striking aspect in [MvT99] is the tentative to perform the search with an imposed group structure: More precisely, it is assumed that there is a subgroup H of the symmetric group S_n such that the elements of the sought-after PSCA can be partitioned in a number of equally sized selections of permutations, each of which corresponds to a

left/right coset[1] of $H \leq S_n$. This incorporation of structure of course potentially prevents the algorithm from finding all types of PSCA$(n, k, 1)$ (or PSCA(n, k, λ), $\lambda \geq 1$ in [Na21]). In particular, the assumption is in general not suitable for showing non-existence results. However, the assumption proves to be extremely useful in both [MvT99] and [Na21]. In the latter, novel exemplars, i.e., $n = 6, 7$, of the type PSCA$(n, 3, 2)$ are herewith calculated, partially settling Yuster's [Yus20] question about the existence of PSCAs with alphabet size $n > 5$ and $(k, \lambda) = (3, 2)$. Notice, that the previously mentioned subgroup H must, in order to allow a packing of permutations into (equally large) cosets, possess a group order being a divisor of $\lambda k!$. Moreover, instead of examining the feasibility of each subgroup $H \leq S_n$ for such a construction, it is enough to check just one representative per each conjugacy class of subgroups of S_n (cf. [Na21]).

A recent work [GW22] performs an exhaustive search by trying to store all PSCAs over n symbols (one representative per equivalence class) and to keep track of all possible PSCAs over $n + 1$ symbols obtainable by inserting the symbol $n + 1$ into the rows. For strength $k = 3$, they reached as extremal case $(\lambda, n) = (3, 8)$, afterwards the search exceeded computation capabilities.

Example 6.2.1. *In [Na21, p. 42], the following representative $\mathscr{A}$ of type* PSCA$(7, 3, 2)$ *is found. It is the union of four right cosets of the subgroup*

$$H := \langle (3, 2, 7, 5, 6, 4, 1) \rangle = \{(1, 2, 3, 4, 5, 6), (3, 2, 7, 5, 6, 4, 1), (7, 2, 1, 6, 4, 5, 3)\} \leq S_7.$$

Consider $\alpha = \mathrm{id} \in S_7$, $\beta = (1, 5, 7, 3, 4, 2, 6)$, $\gamma = (4, 2, 6, 1, 7, 3, 5)$, and $\delta = (4, 7, 5, 6, 1, 2, 3)$ which induce the four right cosets of H,

$$\begin{aligned} H\alpha &= \{(1, 2, 3, 4, 5, 6, 7), (3, 2, 7, 5, 6, 4, 1), (7, 2, 1, 6, 4, 5, 3)\}, \\ H\beta &= \{(1, 5, 7, 3, 4, 2, 6), (3, 6, 1, 7, 5, 2, 4), (7, 4, 3, 1, 6, 2, 5)\}, \\ H\gamma &= \{(4, 2, 6, 1, 7, 3, 5), (5, 2, 4, 3, 1, 7, 6), (6, 2, 5, 7, 3, 1, 4)\}, \\ H\delta &= \{(4, 7, 5, 6, 1, 2, 3), (5, 1, 6, 4, 3, 2, 7), (6, 3, 4, 5, 7, 2, 1)\}. \end{aligned}$$

We set $\mathscr{A} := H\alpha \cup H\beta \cup H\gamma \cup H\delta$ and observe that every sequence in $S_{7,3}$ is covered twice.

6.3 Feasible embeddings of m-sequences

Let us denote the set of all ordered partitions of the number i consisting of j non-negative integer parts by $\mathcal{N}_j(i) \in \mathbb{N}^j$, e.g., $\mathcal{N}_2(4) = \{(0, 4), (4, 0), (1, 3), (3, 1), (2, 2)\}$.

In the following, we study a certain collective behavior resulting from insertion of $k - m$ symbols into a sequence $s \in S_{n,m}$ considered as base case to obtain tuples in $S_{n,k}$ (mainly $m \leq k - 1$ will be of interest; nevertheless we pose the following definition generally.

[1] The selections are either *all* left cosets of H or *all* right cosets of H. We clarify that the left H-coset represented by a is given by $aH := \{a(h(\cdot)) : h \in H\} \leq S_n$. In contrast, the right H-coset represented by a is given by $Ha := \{h(a(\cdot)) : h \in H\}$.

Definition 6.3.1 (Gap-lengths describing vector)**.** *For a permutation $\pi \in S_n$ containing $s = (s_1, \ldots, s_m) \in S_{n,m}$ as subsequence, we consider, for $j = 1, \ldots, m$, the count of symbols in π occurring between the appearances of s_j, s_{j+1}. Collecting all such "gap lengths" we obtain the* gap-lengths describing vector *given by*

$$\mathrm{gdv}_j^{[s,\pi]} := \begin{cases} \pi^{-1}(s_{j+1}) - \pi^{-1}(s_j) - 1, & j = 1, \ldots, m-1 \\ \pi^{-1}(s_1) - 1, & j = 0 \\ n - \pi^{-1}(s_m), & j = m \end{cases}.$$

By its nature, $\mathrm{gdv}^{[s,\pi]}$ is a member of $\mathcal{N}_{m+1}(n-m)$.

Example 6.3.2. *Let $s = (4,1)$, $\pi = (2, \underline{4}, \underline{1}, 3, 5) \approx$ "*4 1**". Then, $\mathrm{gdv}^{[s,\pi]} = (1,0,2)$.*

Definition 6.3.3 (Bag of gap-length describing vectors, BGDV)**.** *Let $n \geq m$ and $\mathscr{A} \subseteq S_n$ be a d-family. Consider $s = (s_1, \ldots, s_m) \in S_{n,m}$ and enlist the permutations of $\mathscr{A}$ containing s as subsequence by $\pi_1, \ldots, \pi_\ell$. We define the* bag of gap-length describing vectors of the pattern s with respect to $\mathscr{A}$*, denoted as* $\mathrm{BGDV}(s, \mathscr{A})$*, as the multiset* $\left\{\mathrm{gdv}^{[s,\pi_i]} : i = 1, \ldots, \ell\right\}$*. Formally,* $\mathrm{BGDV}(s, \mathscr{A})$ *associates to each $u \in \mathcal{N}_{m+1}(n-m)$ a multiplicity $\mu(u, \mathrm{BGDV}(s, \mathscr{A})) \in \{0, \ldots, d\}$.*

Lemma 6.3.4. *Let $n \geq k$ and $m \leq k-1$. Assume that $\mathscr{A} \in \mathrm{PSCA}(n,k,\lambda)$ and that $s = (s_1, \ldots, s_m) \in S_{n,m}$ is fixed. Then,* $\mathrm{BGDV}(s, \mathscr{A})$ *satisfies the equations*

$$\sum_{u \in \mathcal{N}_{m+1}(n-m)} \mu(u, \mathrm{BGDV}(s, \mathscr{A})) \prod_{\ell=1}^{m+1} \binom{u_\ell}{v_\ell} = \lambda \frac{(n-m)!}{(n-k)!} \quad \forall v \in \mathcal{N}_{m+1}(k-m). \tag{6.6}$$

Proof. In string notation,[2] each subsequence $w \in S_{n,k}$ containing s as subsequence can be written as $w = \{*\}^{v_1} s_1 \{*\}^{v_2} \cdots \{*\}^{v_m} s_m \{*\}^{v_{m+1}}$, for an appropriate choice $v = (v_1, \ldots, v_{m+1}) \in \mathcal{N}_{m+1}(k-m)$. The sum in (6.6) does nothing else than counting, for a fixed such "type" v, the supersequences of s of that type covered by $\mathscr{A}$. It is easy to see that in $S_{n,k}$ there are precisely $(n-m)(n-m-1)\cdots(n-k+1) = (n-m)!/(n-k)!$ such "type-v supersequences" being covered λ times by $\mathscr{A}$, explaining the right hand side of (6.6). The equality is fulfilled for an arbitrary partition $v \in \mathcal{N}_{m+1}(k-m)$ such that we arrive at a large number of equations, whose satisfaction is necessary for being a PSCA. □

Assume we want to search for PSCAs $\mathscr{A} \subseteq S_n$ counting $\lambda k!$ permutations. In principle, we can make use of the previous lemma: Pre-calculate (without loss of generality) for the tuple $s = (1, \ldots, m) \in S_{n,k}$ all solutions for $\mathrm{BGDV}(s, \mathscr{A})$ of (6.6), and store these in a collection $\mathcal{B}$. Now, the design of a backtracking algorithm using the information of $\mathcal{B}$ is conceivable. In principle it works as follows: It adds one permutation at the

[2] We use $*$ as wildcard symbol and $\{*\}^\ell$ for denoting the string composed of ℓ wildcards.

time and prunes branches of the search tree immediately when already the conclusion $\mathrm{BGDV}(s, \mathscr{A}) \notin \mathcal{B}$ can be verified for a single $(s_1, \ldots, s_m) \in S_{n,m}$ (in general, this will happen before the λ permutations, that cover s, have been added to the current branch of recursion).

However, an efficient implementation of such an algorithm poses a lot of difficulties and requires a sophisticated memory management (we expect a huge set $\mathcal{B}$). Furthermore, it remains questionable if it really could improve on algorithms of simpler design (cf. the backtracking algorithm in [Na21, NJL22]). For the rest of the section, we limit ourselves to use this concept to force a simplification of the problem, i.e., incorporation of an invariance assumption. The assumption is inspired by the search via cosets [Na21, NJL22] where the enforcement of more structure demonstrates to be a successful speed up for the search.

We analyze the special case that $\mathrm{BGDV}(s, \mathscr{A})$ is invariant under permutations of s, i.e.,

$$\mathrm{BGDV}(s, \mathscr{A}) = \mathrm{BGDV}((s_{\psi(1)}, \ldots, s_{\psi(m)}), \mathscr{A}) \quad \forall \psi \in S_m. \tag{6.7}$$

We will see, that this requirement permits to search for PSCAs only in case they satisfy particular constraints concerning their dimensions. Examples with small $n = 5$, $k = 4$, and $m = 3$ let us suspect that (6.7) is an unnaturally strong requirement for $m > 2$ (probably enforcing the entire symmetric group) – therefore we restrict ourselves to the case $m = 2$.

Lemma 6.3.5. *Let $n \geq k \geq 3$ and let $A \in [n]^{\lambda k! \times n}$ represent $\mathscr{A} \in \mathrm{PSCA}(n, k, \lambda)$. Assume that the gap-length describing vector for any $s = (s_1, s_2)$ is invariant under the flipping $\psi : (s_1, s_2) \mapsto (s_2, s_1)$ (i.e., (6.7) is satisfied for $m = 2$). Then, each column of A has uniform distribution of the symbols in $[n]$ (cf. also Definition 6.4.1). Moreover, it follows that $\lambda k!$ is a divisor of n.*

Proof. When the $\mathrm{BGDV}(\cdot, \mathscr{A})$-values for (s_1, s_2) and (s_2, s_1) are coincident, we have that if the number of rows of A containing (s_1, s_2) as subsequence at index positions (p_1, p_2), $(p_1 < p_2)$, is given by ℓ, then the number of rows containing the reversed subsequence again at positions (p_1, p_2) is given by ℓ, too. Now, consider the first column of A. If a symbol in $[n]$ is not contained in that column, then it cannot be contained in any column: In fact, if it was contained in a such, by reversion it would be present already in the first. The reversion also ensures that the distribution of symbols of the first column is "transferred" to all successive columns. Hence, if one symbol is underrepresented it will be as well in the entire matrix A, contradicting the property of A to contain every symbol equally often. Consequently, also the divisibility condition ensues. □

Note that the property assumed in the previous lemma can be paraphrased as follows.

Definition 6.3.6 (Reflection symmetry)**.** *For n being a divisor of d and $m \leq n$, call a matrix $A \in [n]^{d \times m}$ with no duplicate symbols per row*[3] reflection symmetric*, and write*

[3] This means that the rows can be regarded as elements of $S_{n,m}$.

$A \in \mathcal{R}_{d,m}$, if the following condition holds: For any column indices $j_1 < j_2$, the (j_1, j_2)-submatrix of A satisfies that if h rows coincide with $(a, b) \in [n]^2$, then also h rows coincide with (b, a). The class of all $R \in \mathcal{R}_{d,m}$, which put their rows in lexicographical order, is denoted as $\mathcal{R}^{\text{lex}}_{d,m}$, and we say that such an element R is lex-ordered.

Remark 6.3.7. *We used this terming because of the following geometric construction: Pick two arbitrary columns from a reflection symmetric matrix A, and draw them as $d \times 2$ table of $2 \cdot d$ squares in the plane. Identify $[n]$ with a n-set of distinct colors, and color the cells according to their entries $a_{i,j}$. Then, there will be a suitable reordering of the table's rows such that the resulting drawing is point symmetric with respect to its barycenter (see Figure 6.1 for an illustration).*

Lex-ordering constitutes a canonical form for the elements from $\mathcal{R}_{d,n}$ and allows to define a bijection to multisets of permutations (which is a common representation of PSCAs, cf. [Yus20, NJL22, GW22]).

In particular, the previous assumption implies that the focus is directed to the search within the class of matrices with columnwise equal distribution of the symbols. Let us abbreviate this class by $\mathcal{E}_{d,n}$ (with $n|d$). What if we reject reflection symmetry and search within this larger class? In fact, for selected values such a search has been conducted by [GW22], where the search for $k = 3$ was inconclusive already when encountering $n = 9$.

Before we continue our discussion on the search of PSCAs within $\mathcal{R}_{d,n} \subseteq \mathcal{E}_{d,n}$, let us add a brief comment on this class $\mathcal{E}_{d,n}$.

Remark 6.3.8. *Let $\theta := d/n \in \mathbb{N}$. First let us remark that $\mathcal{E}_{d,n}$ corresponds to the class of* exact $(p, q, x; n)$-Latin rectangles *[AH80], for the choice of parameters $x = 1$ (single-valued matrix entries), $q = 1$ and $p = \theta$ (p, q determine the vertical, respectively horizontal, frequency per symbol). These matrices allow multi-valued matrix entries for $x > 1$; for $(x, p, q) = (1, 1, 1)$ they correspond to classical $n \times n$ Latin squares (cf. [vLW01] and references therein). Note that the class $\mathcal{E}_{d,n}$ contains at least all matrices resulting from interweaving the rows of θ exemplars of $n \times n$ Latin squares. The growth of the count $L(n)$ of $n \times n$ Latin squares has been subject to many studies[4] (cf. e.g. [vLW01]) and is known to be asymptotically equal to*

$$L(n) = \left((1 + o_n(1))\frac{n}{e^2}\right)^{n^2}.$$

This suggests that already searching for PSCAs being the composition of collocated Latin squares might become extremely costly.

In the following, in Algorithm 3, we show how to search recursively for the elements in $\mathcal{R}_{2n,n}$. For this particular choice of parameters, we can come up with a concise

[4] However, no closed formula for the count of Latin squares is available for now. The precise count is known only for $n \leq 11$ (see [SI22a]).

$$\begin{bmatrix}
1&2&3&4\\
1&3&4&2\\
1&4&3&2\\
2&1&4&3\\
2&3&4&1\\
2&4&3&1\\
3&1&2&4\\
3&2&1&4\\
3&4&1&2\\
4&1&2&3\\
4&2&1&3\\
4&3&2&1
\end{bmatrix}
\qquad
\begin{bmatrix}
1&2&3&4&5&6\\
1&4&3&2&6&5\\
2&1&6&5&4&3\\
2&5&6&1&3&4\\
3&5&1&6&2&4\\
3&6&1&5&4&2\\
4&1&5&6&3&2\\
4&6&5&1&2&3\\
5&2&4&3&6&1\\
5&3&4&2&1&6\\
6&3&2&4&5&1\\
6&4&2&3&1&5
\end{bmatrix}
\qquad
\begin{array}{cccccc}
1&2&3&4&5&6\\
1&4&3&2&6&5\\
2&5&6&1&3&4\\
3&5&1&6&2&4\\
3&6&1&5&4&2\\
4&6&5&1&2&3\\
6&4&2&3&1&5\\
6&3&2&4&5&1\\
5&3&4&2&1&6\\
5&2&4&3&6&1\\
4&1&5&6&3&2\\
2&1&6&5&4&3
\end{array}$$

Figure 6.1: Representative of a reflection symmetric PSCA$(4,3,2)$ (left), respectively of a reflection symmetric PSCA$(6,3,2)$ (right). The latter PSCA is printed twice, where the second print illustrates the reflection symmetry with colors for the $(1,2)$-submatrix – after a suitable reordering of rows (see Remark 6.3.7).

n	k	λ	$\#\,\mathrm{PSCA}(n,k,\lambda)\cap\mathcal{R}^{\mathrm{lex}}_{\lambda k!,n}$
3	3	2	1
6	3	2	26 (after 0.7 secs of computation)
9	3	3	? (0 after 20h of computation)
4	4	1	1
6	4	1	≥ 1 (cf. [Na21, Proposition 2.22])
8	4	2	0 (by $g(8,4)>2$, cf. [GW22])
8	4	3	?

Table 6.1: Count of reflection symmetric PSCAs for different parameter constellations. The count/non-emptiness is unknown for $(n,k,\lambda)\in\{(9,3,3),(8,4,3)\}$. The experiment was conducted for $k=3$ via a C++ implementation of Algorithm 3 on a laptop equipped with a AMD Ryzen 5 5500U and 16 gigabytes of RAM running Ubuntu 22.04. Successively duplicates with respect to lex-ordering were removed. The result for $(n,k,\lambda)=(6,4,1)$ follows from the fact that there are only 2 non-equivalent PSCAs [MvT99], one of which is in $\mathcal{R}^{\mathrm{lex}}_{24,6}$.

algorithmic formulation. We note, however, that for every general constellation of parameters d, n, k, λ (with $d = \lambda k!$ and $n|d$), the search within the class $\mathcal{R}^{\text{lex}}_{\lambda k!,n}$ can be tackled (up to computational limits) by a more flexibly designed backtracking algorithm. In Table 6.1 we inspect the count of PSCAs satisfying reflection symmetry. Finding a direct construction for such PSCAs, or a strategy to deny existence of such seems challenging.

Remark 6.3.9. *Even if $n \nmid \lambda k!$, sometimes it is possible to slightly increase n until divisibility holds. It could then be tried to perform the search for this enlarged alphabet size, and if solutions are found, we get PSCAs for the original value n by dropping each symbol not contained in $[n]$. Divisibility can alternatively be ensured by increasing λ.*

We now show how to seek for reflection symmetric PSCAs in $\mathcal{R}_{2n,n}$ via systematic backtracking.

Algorithm 3 Horizontally searching reflection symmetric instances of PSCA$(n, k, 2n/k!)$

Input: Alphabet size $n \in \mathbb{N}^{\times}$, strength $k \geq 3$, multiplicity λ such that $\lambda k!/n = 2$
Output: (Empty) list of members in $\mathcal{R}_{2n,n} \cap \text{PSCA}(n, k, 2n/k!)$

1: **procedure** HorizontalReflectionConstruction(n)
2: Initialize $2n \times n$ matrix $A = (a_{ij})_{ij}$ with blank entries ($2n = \lambda k!$).
3: Populate first row with $(1, 2 \ldots, n)$
4: Populate first column with $c^* := (1, 1, 2, 2, \ldots, n, n)$.

▷ Form the sets of "allowed positions" Q_i per symbol i, telling in which rows i can potentially be inserted:
5: Let $Q_i \leftarrow \{\ell \in [\lambda k!] : a_{\ell,1} \neq i\} \setminus \{1\}$ for each $i = 1, \ldots, n$.

▷ Update dictionary of coverage:
6: Let $C[x] := 1$ for $x \in \{(i_1, \ldots, i_k) : 1 \leq i_1 < i_2 < \ldots < i_k \leq n\}$.

7: Recurse(A, $t = 2$, $(Q_i)_{i=1}^n$, C)

8: **procedure** Recurse(A, t, $\mathcal{Q}$, C)
9: ▷ A: $2n \times n$ matrix $(a_{ij})_{ij}$ having no blank entries within the first $t-1$ columns
10: ▷ t: depth of recursion
11: ▷ $\mathcal{Q}$: A n-tuple of subsets in $2^{[2n]}$ – one subset per symbol in $[n]$
12: ▷ C: Dictionary updating for each $x \in S_{n,k}$ the multiplicity of coverage

13: ▷ Call $u \in [2n]^n$ *row-partitioning* if it selects n row indices $u_1, \ldots, u_n$
14: guaranteeing $(a_{u_1,1}, \ldots, a_{u_n,1}) = (1, \ldots, n)$
15: $U \leftarrow \{u \in [2n]^n : u \text{ is row-partitioning}\}$

16: **for** $u \in U$ **do**
17: $V \leftarrow \left\{r = (r_1, \ldots, r_n) \in S_n : r_1 = t = a_{1,t} \text{ and } u_j \in Q_{r_j}\right\}$
18: **for** $r \in V$ **do**
19: $\bar{r} \leftarrow$ CompleteColumn(r, u, t, A)
20: $\tilde{A} \leftarrow (\vec{a}_1 | \ldots | \vec{a}_{t-1} | \bar{r}) \in [n]^{2n \times t}$, where $\vec{a}_j = (a_{i,j})_{i=1}^n \in [n]^{2n \times 1}$
21: $\tilde{C} \leftarrow C$ ▷ Make copy of coverage information

22: **if** $t \geq k-1$ **then**
23: Update in $\tilde{C}$ all yet-untracked coverages observable in $\tilde{A}$

24: **if** $\bigwedge_{j=2}^{t-1} [(\vec{a}_j | \bar{r}) \in \mathcal{R}_{2n,2}]$ **then**
25: **if** $t = n$ **then**
26: **if** $\tilde{C}$ attests perfectness **then**
27: $L \leftarrow L \cup \left\{\tilde{A}\right\}$ ▷ PSCA found
28: **else**
29: ▷ Check for excess coverage
30: **if** $\nexists x : \tilde{C}[x] > \lambda$ **then**
31: $\tilde{Q}_i \leftarrow Q_i \setminus \{j \in [2n] : \bar{r}_j = i\}$ for each $i = 1, \ldots, n$.
32: Recurse($\tilde{A}$, $t+1$, $(\tilde{Q}_i)_{i=1}^n$, $\tilde{C}$)
33: **return** L

34: **procedure** CompleteColumn(r, u, t, A)
35: Find the unique vector $\bar{r} \in [2n]^n$, whose entries
– constitute a reordering of c^*,
– extend r, i.e., $\bar{r}_{u_1} = r_1, \ldots, \bar{r}_{u_n} = r_n$, and
– guarantee that the $(2n) \times 2$ matrix $((a_{u(i),t})_{i \in [n]} | \bar{r}) \in \mathcal{R}_{2n,2}$.
36: **return** $\bar{r}$

6.4 General search with pre-determined distributions

Columnwise distribution of symbols for PSCAs has been analyzed very recently in [GW22]. In the following we sketch a search based on an assumed columnwise distribution of symbols (of arbitrary shape). Afterwards, we remark some observations due to [GW22].

Setting $m = 1$ in Lemma 6.3.4 we obtain as special case the following Lemma 6.4.2 already observed in [GW22].

Definition 6.4.1. *Let $\mathscr{A} = \{\pi_1, \ldots, \pi_r\} \subseteq S_n$. For a symbol $x \in [n]$, the* (columnwise) distribution vector *of x, denoted $d^{[x]}$, is defined by*

$$d_j^{[x]} := |\{i \in [r] : \pi_i(j) = x\}|, \ j = 1, \ldots, n.$$

Lemma 6.4.2. *If $\mathscr{A} = \{\pi_1, \ldots, \pi_{\lambda k!}\} \subseteq S_n$ lies in* PSCA(n, k, λ)*, then for each symbol $x \in [n]$, the distribution vector must satisfy the equations*

$$\sum_{j=1}^{n} d_j^{[x]} \binom{j}{\kappa} \binom{n-1-j}{k-1-\kappa} = \lambda \frac{(n-1)!}{(n-k)!} \quad \text{for all } \kappa = 0, \ldots, k-1, \tag{6.8}$$

$$\sum_{j=1}^{n} d_j^{[x]} = \lambda k!, \text{ with } d^{[x]} \in \{0, 1, \ldots, \lambda k!\}^n . \tag{6.9}$$

□

Remark 6.4.3. *If $(\nu_1, \ldots, \nu_n) \in \mathbb{N}^n$ is a solution for $d^{[x]}$ in* (6.8)*, then it is as well its reversal $(\nu_n, \nu_{n-1}, \ldots, \nu_1)$. This behavior is expectable due to the invariance of PSCAs under reversion of all rows (cf. Lemma 2.3.8).*

Remark 6.4.4 (A general horizontal construction approach)**.** *If for any symbol $x \in [n]$ its frequency per column is fixed, this leads to an idea on how to seek horizontally, i.e., column by column, for a PSCA: The first column can be an arbitrary (e.g. lexicographically ordered) vector respecting the pre-imposed frequencies of symbols. The second column can be sampled according to its associated distribution, but has furthermore to satisfy the condition that its entries differ from their left predecessors. In a backtracking approach columns are recursively added with the goal to reach, in principle, all the distribution respecting matrices $[n]^{\lambda k! \times n}$ corresponding to the leaves in the search tree. The more columns are already present (depth of the backtracking), the less possibilities (if even any exist) for the successive column are available. Moreover, to ensure that perfect coverage is never violated during the transition of the search tree, the information of covered k-sequences in $S_{n,k}$ should be carried over and updated, and branches leading to excess coverage ($> \lambda$) should be immediately pruned.*

The following lemma shows that determining how to combine distributions is again a "packing problem" describable by a linear equation system to be solved over a high dimensional orthant of a Cartesian power of $\mathbb{N}$.

Lemma 6.4.5. *Let $\mathscr{A}$ be a* PSCA(n, k, λ) *and let $D \subseteq \mathbb{N}^n$, enumerated by $d_1, \ldots, d_{|D|}$, contain the columnwise distributions obtainable from solving the corresponding system* (6.8)-(6.9)*. Consider $\gamma_1, \ldots, \gamma_{|D|} \in \mathbb{N}$, where γ_i stands for the count of symbols in $[n]$ following the column-distribution d_i. Then,*

$$\sum_{i=1}^{|D|} \gamma_i d_i = \lambda k! \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \left(\mathbb{N}^{\times}\right)^n, \tag{6.10}$$

$$\sum_{i=1}^{|D|} \gamma_i = n. \tag{6.11}$$

□

Remark 6.4.6. *We notice that finding (all) solutions for* (6.10)-(6.11) *could be achieved by finding (all) integral points in the convex polytope formed by the points of $\mathbb{R}^n$ satisfying each of the involved equations/inequalities. Alternatively, a systematic backtracking approach seems conceivable. Also making use of a CSP[5] solver could be a viable option. In the literature (cf. [HLS10]), we found a hard, closely related (optimization) problem appertaining to the class of knapsack problems: It is known as* multidimensional multiple choice knapsack problem*.*

Example 6.4.7. *Let $n = 5$, $k = 3$, and $\lambda = 2$. Searching all points in $\mathbb{N}^n$ satisfying the system* (6.8)-(6.9)*, we obtain that there are only* 8 *possible columnwise distributions for any symbol $x \in [n]$, more precisely we have*

$$d^{[x]} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \end{pmatrix} \in \left\{ \begin{pmatrix} 0 \\ 8 \\ 0 \\ 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \\ 0 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \\ 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 3 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 0 \\ 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 3 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 6 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 0 \\ 0 \\ 8 \\ 0 \end{pmatrix} \right\} =: D. \tag{6.12}$$

We conducted the search by passing the respective equation system via Python to the solver OrTools[6] (version 9.4.1874) able to handle integer linear problems among others. The size of the solution set is in line with the results in [GW22].

We continue on Example 6.4.7.

Example 6.4.8. *Let $d_1, \ldots, d_8$ enlist the vectors in* (6.12)*. It turns out that the solution set of the corresponding system* (6.10)-(6.11) *is given by the following* 17 *vectors in*

[5] Constraint Satisfiability Program.

[6] `https://pypi.org/project/ortools/`

$\mathbb{N}^8$ *(displayed as columns). Again the results are provided by the solver OrTools (cf. Example 6.4.7).*

0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	1	1	1	0	1	2	0	0	0	0	0
1	2	3	0	1	0	1	0	1	0	0	0	0	0	0	0	1
2	1	0	3	2	1	0	1	0	4	2	0	0	0	0	1	0
0	0	0	1	1	0	0	1	1	0	0	0	0	1	2	0	0
2	1	0	1	0	3	2	1	0	0	0	0	4	2	0	1	0
0	1	2	0	1	0	1	1	2	0	1	2	0	1	2	1	2
0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1

Example 6.4.9. *A C++ prototype implementation of the backtracking procedure sketched in Remark 6.4.4 fed with the legit combinations of distribution vectors from Examples 6.4.7 and 6.4.8 yields that a* $\mathrm{PSCA}(5,3,2)$ *exists for the combination vectors in the table of Example 6.4.8 with column indices in* $\{1,2,3,10,13,16\} \subseteq [17]$*. We display the very first PSCA found with this backtracking approach (hereby, we have* $d^{[1]} = d_3$*,* $d^{[2]} = d^{[3]} = d_4$*, and* $d^{[4]} = d^{[5]} = d_6$*):*

$$\begin{bmatrix}
1 & 2 & 4 & 5 & 3 \\
1 & 3 & 4 & 5 & 2 \\
2 & 1 & 5 & 4 & 3 \\
2 & 3 & 4 & 1 & 5 \\
3 & 1 & 5 & 4 & 2 \\
3 & 2 & 5 & 1 & 4 \\
4 & 1 & 3 & 2 & 5 \\
4 & 2 & 3 & 5 & 1 \\
4 & 5 & 2 & 1 & 3 \\
5 & 1 & 2 & 3 & 4 \\
5 & 3 & 2 & 4 & 1 \\
5 & 4 & 3 & 1 & 2
\end{bmatrix}$$

In some circumstances, the search of legit columnwise distributions can be accomplished by solving linear systems of congruences (see (6.13)). The derivation of such a system of congruences is based on the following more general result.

Theorem 6.4.10 ([GW22])**.** *Let* $n \geq k \geq 2$*,* $\lambda \geq 1$*, and* $\mathscr{A} \in \mathrm{PSCA}(n,k,\lambda)$ *with columnwise distributions of symbols* $d^{[\cdot]}$*. For any symbol* $x \in [n]$ *and any* $\kappa \in [k-1]$*, we have*

$$\frac{1}{\lambda k!} \sum_{j=0}^{n-1} j^\kappa d_{j+1}^{[x]} = \frac{1}{n} \sum_{i=0}^{n-1} i^\kappa.$$

□

Theorem 6.4.11 ([GW22])**.** *Let $n \geq p \geq 3$ for a prime p. Suppose that $n \not\equiv 0 \pmod{p}$. Consider $\mathscr{A} \in \mathrm{PSCA}(n, p, \lambda)$ with columnwise distributions of symbols $d^{[\cdot]}$. Moreover, let $x \in [n]$ be a fixed symbol and $r \in \{0, \ldots, p-1\}$ a residue. Then,*

$$\sum_{j \in \{0,\ldots,n-1\}:\ j \equiv r \pmod{p}} d_{j+1}^{[x]} \equiv 0 \pmod{p}. \tag{6.13}$$

□

Moreover, in [GW22, Theorem 2.7] a strategy to rule out some columnwise distributions for $\mathrm{PSCA}(n, k, \lambda)$ by exploiting an interconnection to distributions for $\mathrm{PSCA}(n-1, k, \lambda)$ is derived. Interestingly, empirical results in [GW22, Table 3] show that the fraction of herewith ruled out distributions seems to converge to zero with increasing n.

CHAPTER 7

Applications

In the following we provide a selection of application fields showing that the analysis and study of scrambling permutations (including many of its analogues and spin-offs) have launched some useful findings of a theoretical and also practical nature.

7.1 Order theory, combinatorial geometry and related areas

When introducing k-scrambling permutations, Dushnik noticed that $N(n,k)$ could be used to describe the dimension $\dim(1,k;n)$ of a specific class of partial orders (1- and k-subsets of $[n]$ are hereby ordered by inclusion, see [Dus50] for details). The concept of dimension referred to here, in fact, was coined earlier by the latter together with Miller in [DM41], and is defined for a partial order P to be the least number of linear extensions of P whose intersection is P. In contrast to Dushnik, Trotter later considered also $\dim(1,k;n)$ when k is small (cf. [Tro78]).

In [HM99], the authors relied on scrambling permutations for analyzing when particular simplicial polytopes form a Scarf complex. In [Für00], Füredi used scrambling permutations to obtain a simple proof for the boundedness of the Prague dimension of Kneser graphs. The number $N(n,k)$ is also characteristic in the context of a convex geometries for the concept of the convex dimension [ES88, BM14]. More recently, in [SW20] again the upper bound (4.2) for $N(n,k)$ is utilized in order to establish an upper bound for the boxicity of a graph depending on the maximal degree of the latter. An alternative bound for the boxicity depending on the Euler genus of the graph is obtained as well via (4.2).

From the motivation to improve bounds for containment problems in high-dimensional space, direct or indirect contributions to the sharpening of bounds for completely scrambling permutations were obtained. Indeed, in [Für96] Füredi improved the bound for

the following problem by exploiting the equivalence to the estimation of $N^*(n,3)$ (the equivalence was already known to Ishigami in [Ish95]).

Problem 7.1.1. *For the Euclidean space $\mathbb{R}^d$ find the maximal number n such that there exists a selection of distinct points $x_1,\ldots,x_n \in \mathbb{R}^d$ fulfilling the following condition: For $i=1,\ldots,n$, inside any possible axis-parallel (topologically closed) orthant anchored at x_i there lies at most one additional point x_j $(j \neq i)$.*

In [BCG+16] the 3-mixing property (recall Definition 2.2.3) is used to relate the separation dimension of a graph to its so-called acyclic chromatic number. Radhakrishnan's idea for bounding $N^*(n,k)$ from below was utilized in the latter work in order to analyze lower bounds for the separation dimension of hypergraphs.

7.2 Combinatorial software testing

In the following we point out a selection of practical aspects of sequence covering arrays for the purpose of *event sequence testing*. We start by giving below a minimalist example scenario useful for automatically finding potential failures in the realization of an operating system (hereby the selected application scenario is fictive and based on analogous considerations illustrated in the literature [KHL+12, BEI+12]).

Example 7.2.1. *Assume the operating system is tested after having successfully booted and that following events could be actuated by a user:*

- `EIUSB` – *Establish internet connection via USB port*
- `CWLAN` – *Connect to a WLAN*
- `CWCAM` – *Connect peripheral device webcam*
- `ENBLU` – *Enable Bluetooth connectivity*
- `CMOUS` – *Connect peripheral device mouse*
- `LOGOI` – *Logout and successively login*
- `SUSWU` – *Suspend and wakeup*

If the operating system as well as all involved components are free of bugs, then executing all of the above seven events (each event only once) in any chronological order should trigger no error/failure of the system. Therefore, for ensuring such a well-behavior a number of tests super-exponential in the count of events (here for our illustration $7! = 5040$) have to be carried out. If each single test is interpreted as one row of a SCA, then this test scenario would correspond to a SCA with alphabet size 7 and strength 7. One might be interested in examining only event sequences corresponding to a SCA

Performed test							Error triggered?
EIUSB	CWLAN	CWCAM	ENBLU	CMOUS	LOGOI	SUSWU	/
ENBLU	CWCAM	CWLAN	EIUSB	SUSWU	LOGOI	CMOUS	/
EIUSB	CMOUS	LOGOI	SUSWU	CWLAN	CWCAM	ENBLU	/
SUSWU	LOGOI	CMOUS	EIUSB	ENBLU	CWCAM	CWLAN	/
CWLAN	CMOUS	EIUSB	CWCAM	ENBLU	LOGOI	SUSWU	/
CMOUS	CWLAN	SUSWU	LOGOI	ENBLU	CWCAM	EIUSB	Yes
CWCAM	LOGOI	EIUSB	CWLAN	ENBLU	CMOUS	SUSWU	/
LOGOI	CWCAM	SUSWU	CMOUS	ENBLU	CWLAN	EIUSB	Yes
ENBLU	SUSWU	EIUSB	CWLAN	CWCAM	CMOUS	LOGOI	/
SUSWU	ENBLU	LOGOI	CMOUS	CWCAM	CWLAN	EIUSB	Yes

Figure 7.1: Usage of a SCA for error detection. Each row corresponds to a test of seven chronologically actuated actions. It is the output of Tarui's construction of Section 5.2 with parameters $(n, k) = (7, 3)$.

with same alphabet size but with lower strength parameter. A motivation for this weaker requirement is that one might perform the testing under the assumption that an error-triggering event does not depend on all but only on a small number of events chronologically happening before. Following this approach, a test suite corresponding to a respective SCA of strength 3 could be enough for detecting failures (e.g. the one displayed in Figure 7.1 which lets suspect that there is a misconfiguration of the system causing a failure occurring when `SUSWU`, `ENBLU`, `EIUSB` appear (chronologically) in succession). Optimally, each test can be executed with the same starting conditions (in our example this could be achieved by testing the system in a virtual machine). From the logarithmic bound for fixed strength (see Proposition 4.2.1), or from Figure 7.1, we can notice a drastic reduction of tests to be carried out. If for the previous example use-case a SCA of strength 4 is employed as test suite, the number of tests can be reduced to 38 (this value is the actually lowest available and obtained in [BTI12]) instead of 5040.

The methodology's impact gets particularly significant if large sets of events are taken in consideration (such that testing every permutation gets impracticable). On the other hand also for small event sets, if performing a test is highly expensive in terms of time or (e.g. logistic or monetary) costs, the method qualifies for being a viable measure.

Indeed, the legitimacy of such assumptions is justifiable by referring to a similar assumption in the closely related field of combinatorial testing with covering arrays: Here, for a selection of application domains, in [KKL12] empirical results providing strong evidence for the plausibility of such assumptions are presented.

It becomes clear that for many applications additional precautions have to be taken into account, e.g., certain undesirable sequences have to be excluded during the generation of a test suite (they could, e.g., damage the system under test, or such a behavior of the system can never occur/is not of interest). In order to fulfill such additional conditions, in [KHL$^+$12] it is outlined how this can be accomplished by semi-supervised expansion

of smaller test suites. A realization of test suites with highly flexible constraints via answer set programming is explained in [BEI+12]. It should be noted that sequences in which the same event occurs (chronologically) several times in succession cannot be handled directly in the setting of SCAs. However, a way out of this is conceivable, e.g. by carrying out the generation of a SCA and retrospectively identifying a certain number of originally different alphabet symbols.

The principles of the above described methodology are reflected e.g. in the design of test suites for testing the graphical user interface of applications [YCM07, MGB+16, ANPB18], or the construction of navigation graphs for dynamic web applications [WLS+09]. Hereby, some of these constructions fall back not exclusively on the concept of SCAs but also take up some useful aspects of CAs, which have been studied for much longer. CAs, in fact, are an even more widespread, popular concept and are applied for detecting erroneous (but also malicious) behavior of software/hardware systems (among the numerous works addressing these topics cf. e.g. [Har05, CFT14, HWKK20]).

In [GHLS93] CAs are also used to quantify the fault tolerance of hypercube computers (a parallel computer possessing 2^n processors and a specific network topology), and in [Har05] they are applied to the so-called "minimax rendezvous time problem for k distinguishable robots on a line".

7.3 Min-wise hashing

This section briefly discusses the applications of min-wise independent permutations. Before we come to them, we discuss the role of ε-approximately min-wise independence and afterwards the notion of so-called linear permutations.

As min-wise independent families are exponentially large (in n), for applications where n is typically of magnitude 2^{64} (cf. [BCFM00]), the trick of using ε-approximately min-wise independence as compromise is applied. In fact, for this approximate variant of min-wise independence families of quadratic size can be found. For ε-approximately k-restricted min-wise independence, even families of logarithmic size are obtainable (see Theorem 4.3.3). Moreover, note that in praxis only approximations of the uniform probability distribution can be achieved computationally (cf. [BCFM00]) such that the study of permutation families endowed with not necessarily uniform probability distributions is of high importance.

We now present a subclass of permutation families which excel in simplicity. Consider the finite field $\mathbb{Z}_p$, for a prime p. Affine maps $\mathsf{p}_{a,b} : \mathbb{Z}_p \to \mathbb{Z}_p$, $x \mapsto ax + b$ ($a \neq 0$) appertain to the symmetric group $S_{\{0,\dots,p-1\}} \subseteq \mathbb{Z}_p^{\mathbb{Z}_p}$. Consequently, by natural identification of $\mathbb{Z}_p$ with $[p]$, the notion of affinity can be lifted to S_p. We define $\mathrm{Aff}_p := \{\pi \in S_p : \pi \text{ is affine}\} \subseteq S_p$ and call its members *linear permutations* [BCFM00]. The quality of linear families (i.e., consisting of linear permutations) are discussed by the inventors of min-wise independent families in [BCFM00]: The following result shows that for large values of p and k, linear

families are k-restricted min-wise independent up to an on average small approximation error.

Theorem 7.3.1 ([BCF00])**.** *Assume that the set of all $X \subseteq [p]$ with $|X| = k$ is endowed with uniform probability distribution. Then, for $p, k \to \infty$, when π is a permutation chosen uniformly at random from Aff_p, we have*

$$\mathop{\mathbb{E}}_{X:|X|=k}\left[\max_{x\in X}\{\Pr[\pi(x) = \min \pi(X)]\}\right] = \frac{1}{k} + O\left(\frac{(\ln k)^3}{k^{3/2}}\right).$$

□

Broder et al. [BCFM00] introduced min-wise independent permutations for the purpose of detecting near-duplicate documents on the internet. The concept is reflected in a practically operable algorithm of the AltaVista web index software (cf. [BCFM00] for further details). Initially, a set is associated to each document by the so-called *w-shingling* technique [Bro97]; here $w \in \mathbb{N}$ is a small value. The main idea is the following: The set of all w-shingles, i.e., all w-tuples consisting of w consecutive words in the document, is stored as document-characterizing set (or multiset – for higher precision). Let us assume that the w-shingles are replaced by natural numbers, such that documents are characterized by subsets of $[n]$ for (large) n. If A and B determine these sets for two documents a and b, the *resemblance*[1] *of A and B* can be used to measure similarity between a and b. It is defined as

$$r(A, B) := \frac{|A \cap B|}{|A \cup B|} \in [0, 1].$$

According to [Bro97] this measure seems to be well suited to represent the notion of "roughly the same", when $r(A, B) \approx 1$. Moreover, $\delta(A, B) := 1 - r(A, B)$ is a metric [Bro97]. It has been noticed in [Bro97] that if π is drawn (with uniform probability) from a min-wise independent family $\mathscr{F}$ of permutations, then

$$\Pr[\min \pi(A) = \min \pi(B)] = r(A, B). \tag{7.1}$$

In [BCFM00] the observation (7.1) is used for introducing, for a document a, the so-called *sketch* ($A \subseteq [n]$ denotes the set of w-shinglings of a) given by

$$\overline{S_A} := (\min \pi_1(A), \min \pi_2(A), \ldots, \min \pi_L(A)), \tag{7.2}$$

where $\pi_1, \ldots, \pi_L$ are L randomly drawn, and henceforth fixed, permutations from $\mathscr{F}$. Given two sketches $\overline{S_A}$ and $\overline{S_B}$ for two documents a and b, the resemblance $r(A, B)$ can be estimated by the fraction of coincident elements of $\overline{S_A}$ and $\overline{S_B}$. In praxis, this principle is used for (ε-approximately) k-restricted min-wise independent families. For a large amount of documents this method of comparison reduces computational costs

[1] Also known as *Jaccard similarity* [LK11].

as only the sketches (small-sized in comparison to the entire document size) have to be compared. For a performance analysis of the Min-wise hashing algorithm (MinHash) on real-world data, we point to a study conducted in [Hen06], where the algorithm is classified as "state-of-the-art" and its relevance for "successful web search engines" is emphasized.

Since their introduction, min-wise independent permutations have found use in several application areas, which include content matching for online advertising, detection of Web spam and redundancy in file systems, etc. (cf. [LK11] and references therein). In [CM10] Min-hashing has been employed for "fast detection of pairs of images with spatial overlap". Further applications of Min-wise hashing address dimensionality reduction for machine learning (cf. e.g. [ZMA16]), derandomization of algorithms [BCM98], and computational geometry [Mul94].

CHAPTER 8

Conclusion

8.1 Discussion

We inspected scrambling permutations and surveyed related structures (some of which turned out to be isomorphic circumscriptions). The encountered combinatorial structures satisfy a number of interacting/concurring conditions, which rapidly grows in the parameters n and k. A priori, this gives the impression that questions about minimal representatives can only be determined brute-force – except for special cases, this is indeed the case according to the current state of knowledge. Nevertheless, so far, approaches have been achieved that allow relatively well to narrow down the orders of magnitude of optimal families (where the quality depends on the type of the family). For many insights, an interplay of diverse mathematical branches is indispensable, among others, combinatorics, probability theory, information theory, coding theory and finite geometry. Moreover, the gained insights have a great impact on a wide variety of applications.

For the derivation of lower bounds it is necessary to find suitable relaxed and more easily manageable structures. On the other hand, (practicable) constructions – and consequently upper bounds – are obtained by derandomization of initially asymptotically formulated results. Here, the probabilistic method is a recurrent tool to find in a first step upper bounds for various structures. According to current knowledge, the bounds found in this way can be improved only rarely and to a small extent. Depending on the concept or special case, we have explained approaches (Hajnal's or Tarui's construction) which, on the other hand, are based on a different principle: a set system is introduced in a clever way, which is then used to specify collections of linear orderings on it which satisfy the required scrambling conditions.

For optimal families of (completely) scrambling permutations, the question of asymptotic growth of their cardinality is essentially solved: It is of order $\Theta(\log_2 \log_2 n)$ (respectively $\Theta(\log_2 n)$). We have pointed out that lower bounds for completely scrambling permuta-

tions obtained by using a relation to binary CAs cannot be used to improve the bounds of Füredi–Radhakrishnan relying on entropy methods.

In contrast, the growth of the corresponding regularized families is confined only up to a certain degree of accuracy (polynomials of different degree for lower and upper bounds of PSCAs/min-wise independent families). For the growth of PSCAs, even for the special case $k = 3$, it is still unclear by which factor the bounds can be tightened: We pointed out that the upper bound for PSCAs given in [Yus20] can be improved by a factor of approximately $(\log_2 n)^{0.81}$ in any case, so that the growth of $g(n, 3)$ is at least linear in n and at most $O(n(\log_2 n)^2)$. For the more general case, we were able to answer an open question of R. Yuster: There always exist PSCAs of polynomial size. Moreover, we have improved $g(n, k)$-values for a selection of explicit parameter constellations (n, k) of small magnitude.

The task of constructing such regularized structures has proven to be particularly challenging (in particular, since one has to stick to exact structure sizes in a dedicated way, i.e. approximative algorithms for the non-regular case do not bear fruit). In this context, the following questions immediately arose: leaving aside exorbitantly large collections of permutations (such as the entire symmetric group), are there any "small" such collections at all? Knowing that lower bounds apply, there is always a jump point for which this optimal packing of subpermutations of S_n succeeds, but not for S_{n+1}; the question of the underlying mechanism causing this impossibility always resonates.

Given the challenging construction of optimal PSCAs for small parameters n, k, and λ, a current trend is to seek/construct PSCAs that have a particular group-theoretic structure (cf. [Na21, NJL22]) leading to not necessarily optimal instances. We have given insights into this trend and finally joined it with the proposal to search PSCAs in a space of matrices with a certain refection symmetry property. The idea for this came from considering a generalization of the columnwise frequency vectors analyzed in [GW22].

8.2 Open questions and further work

Problem 8.2.1 (cf. [BEI$^+$12, CCHZ13])**.** *Establish the right complexity class of the problem to calculate the exact value of $N^*(n, k)$ and to calculate the exact value of $g(n, k)$.*

Problem 8.2.2. *Verify/falsify the revisited Levenshtein conjecture (see [MvT99]), stating that for $k \notin \{1, 2, 4\}$, the estimate $g(k + 2, k) > 1$ applies. It is known to be true for $k \in \{3, 5, 6\}$ and is unresolved for $k \geq 7$.*

Problem 8.2.3 (cf. [Tar08])**.** *Find a direct construction (comparable to the one of Tarui) to generate completely k-scrambling families of small size, for $k = 4$ or for other values $k \geq 4$.*

The following problem is due to C. Colbourn [NJL22] .

Problem 8.2.4. *Instead of $g(n,k)$, the minimum value λ for which* $\mathrm{PSCA}(n,k,\lambda)$ *is non-empty, find the minimum value λ^* for which* $\mathrm{PSCA}(n,k,\lambda)$ *is nonempty for all $\lambda \geq \lambda^*$.*

Problem 8.2.5 ([GW22, NJL22])**.** *Determine the exact value of $g(9,3) \in \{3,4\}$.*

Problem 8.2.6. *Verify/falsify the following conjecture: $g(n,3) = \Theta(n \log_2 n)$.*

Our observations in Section 6.3 raise the following questions: Under the assumption $n|\lambda k!$, it would be interesting to answer the following question (or to come up with explicit counterexamples): Does existence of $\mathscr{A} \in \mathrm{PSCA}(n,k,\lambda)$ automatically imply existence of an exemplar $\bar{\mathscr{A}} \in \mathrm{PSCA}(n,k,\lambda) \cap \mathcal{E}_{\lambda k!,n}$, i.e., such that $\bar{\mathscr{A}}$ has columnwise uniform distribution of symbols? If the answer is affirmative this leads to another question: Could it be that the PSCAs of class $\mathrm{PSCA}(n,k,\lambda) \cap \mathcal{E}_{\lambda k!,n}$ are the ones that are the "easiest" to find? We proposed to search PSCAs within $\mathrm{PSCA}(n,k,\lambda) \cap \mathcal{R}_{\lambda k!,n}$ and hit computational limitations already for $k = 3$ and $n = 9$. Another question concerns the rarity of the class $\mathcal{R}_{\lambda k!,n}$ itself: Is it possible to find bounds/exact numbers for the count of $(\lambda k!) \times n$ matrices in that class?

For small values of n, it would be interesting to analyze if k-restricted min-wise independent families can be efficiently found by systematically searching within unions of cosets of S_n (in Example 4.3.10 we encountered a manually constructed representative of such a nature). Such an analysis would lift the approaches pursued in [Na21, NJL22] to a more general (up to isomorphic identification) combinatorial structure and could provide some insights about the right order of magnitude of cardinalities of optimal k-restricted min-wise independent families.

The concept of approximateness for min-wise independent families could be lifted to PSCAs (or rankwise independent permutations). In sense we could term this property as ε-perfectness: For fixed $k \in \mathbb{N}$, a d-family $\mathscr{A} \subseteq S_n$ fulfills ε-perfectness iff for a permutation π drawn uniformly at random from $\mathscr{A}$, we have

$$\left| \Pr\left[\pi \text{ covers } (x_1,\ldots,x_k)\right] - \frac{1}{k!} \right| \leq \frac{\varepsilon}{k!}, \text{ for each } (x_1,\ldots,x_k) \in S_{n,k}.$$

Denoting by $\varphi_{(x_1,\ldots,x_k)}$ the count (with multiplicity) of permutations in $\mathscr{A}$ which cover $(x_1,\ldots,x_k)$, the latter relative error means that for all $(x_1,\ldots,x_k) \in S_{n,k}$,

$$\left\lceil \frac{d}{k!} - \frac{\varepsilon d}{k!} \right\rceil \leq \varphi_{(x_1,\ldots,x_k)} \leq \left\lfloor \frac{d}{k!} + \frac{\varepsilon d}{k!} \right\rfloor. \tag{8.1}$$

As already for the question whether there is a PSCA for given strength k and fixed number of rows d, one could ask the following (now, d is not necessarily a multiple of $k!$): If we additionally fix a relative error tolerance ε, is there a d-family $\mathscr{A} \subseteq S_n$ satisfying (8.1) for the frequencies $\varphi_{(x_1,\ldots,x_k)}$ for all $(x_1,\ldots,x_k) \in S_{n,k}$? It would be interesting to (computationally) determine/estimate for given d and k the smallest possible value

for ε still satisfying the bound (8.1). This might be helpful to get better insights why some parameter constellations of strength and multiplicity do not permit existence of PSCAs. We notice that instead of keeping small in (8.1) the margin of $\varphi_{(x_1,\ldots,x_k)}$ with respect to the maximum norm, it might also be interesting how small the margin can be tightened in terms of other metrics such as the L^1 norm. An easily noticeable property of PSCAs is that, among all d-families of permutations, they minimize the cardinality of $\text{range}(\varphi) = \left\{\varphi_{(x_1,\ldots,x_j)} : (x_1,\ldots,x_k) \in S_{n,k}\right\}$ (as the latter coincides with the singleton $\{\lambda\}$ for PSCAs). To get an alternative weakened concept of perfectness, in contrast to keeping the margin of error tolerance small (like in (8.1)), we could require the minimization of the cardinality of $\text{range}(\varphi)$. Having components with a small palette of frequencies, could perhaps make it more manageable to combine the components to a larger, ordinary PSCA.

We keep the arisen questions for further investigation.

List of Figures

2.1 Venn diagram classifying the various properties of permutation families (with fixed n and k, both of general character). 17

3.1 The codes $\mathrm{VT}^{4,a}$, $a = 0, \ldots, 4$ partitioning $\{0,1\}^4$. 24

4.1 A comparison of lower bound coefficients $c(k) \in \{c_{\mathrm{FR}}(k), c_{\mathrm{KS}}(k)\}$ (for small values of k) asymptotically satisfying $N^*(n,k) \geq c(k) \log_2 n$. 42

6.1 Representative of a reflection symmetric PSCA$(4,3,2)$ (left), respectively of a reflection symmetric PSCA$(6,3,2)$ (right). The latter PSCA is printed twice, where the second print illustrates the reflection symmetry with colors for the $(1,2)$-submatrix – after a suitable reordering of rows (see Remark 6.3.7). . 66

7.1 Usage of a SCA for error detection. Each row corresponds to a test of seven chronologically actuated actions. It is the output of Tarui's construction of Section 5.2 with parameters $(n,k) = (7,3)$. 75

List of Tables

2.1 Explicit verification of the property of being 3-mixing for $\mathscr{P} = (\pi_1, \ldots, \pi_5)$ of Example 2.2.4. 8
2.2 Explicit verification of the property of being completely 3-scrambling for $\mathscr{Q} = (\pi_1, \ldots, \pi_6)$ of Example 2.2.4. 8

4.1 Upper bounds for $g(n, 4)$ (cf. [Iur22]). Entries beating best-known values are in bold print. 49
4.2 Upper bounds for $g(n, 3)$. Entries beating best-known values are in bold print. 49
4.3 Summary of bounds encountered in this chapter. Dashes indicate that the more general bound (arbitrary k) should be consulted. Coefficients in bold print highlight improvements pointed out. The results stated for PSCAs are valid for rankwise independent families, too. 50

6.1 Count of reflection symmetric PSCAs for different parameter constellations. The count/non-emptiness is unknown for $(n, k, \lambda) \in \{(9, 3, 3), (8, 4, 3)\}$. The experiment was conducted for $k = 3$ via a C++ implementation of Algorithm 3 on a laptop equipped with a AMD Ryzen 5 5500U and 16 gigabytes of RAM running Ubuntu 22.04. Successively duplicates with respect to lex-ordering were removed. The result for $(n, k, \lambda) = (6, 4, 1)$ follows from the fact that there are only 2 non-equivalent PSCAs [MvT99], one of which is in $\mathscr{R}^{\text{lex}}_{24,6}$. 66

List of Algorithms

1 Deterministically keeping Spencer's bound 55
2 Generating 3-scrambling permutations via Tarui's construction 57
3 Horizontally searching reflection symmetric instances of PSCA$(n, k, 2n/k!)$ 67

Bibliography

[AH80] Lars Døvling Andersen and Anthony JW Hilton. Generalized Latin rectangles ii: embedding. *Discrete Mathematics*, 31(3):235–260, 1980.

[Alo03] Noga Alon. Problems and results in extremal combinatorics – i. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[And87] Ian Anderson. *Combinatorics of Finite Sets*. Oxford science publications. Clarendon Press, 1987.

[ANPB18] David Adamo, Dmitry Nurmuradov, Shraddha Piparia, and Renée Bryce. Combinatorial-based event sequence testing of Android applications. *Information and Software Technology*, 99:98–117, 2018.

[AS00] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience series in discrete mathematics and optimization. Interscience Publishers, New York, second edition, 2000.

[Bar04] Victor Bargachev. An improved lower bound on the size of k-rankwise independent families of permutations, 2004. Preprint on website of St. Petersburg Department of Steklov Institute of Mathematics [online] `https://www.pdmi.ras.ru/preprint/2004/04-13.html`(last access: 2022-08-17).

[BCF00] Tom Bohman, Colin Cooper, and Alan Frieze. Min-wise independent linear permutations. *The electronic journal of combinatorics*, 7(1):R26, 2000.

[BCFM00] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, 2000.

[BCG+16] Manu Basavaraju, L. Sunil Chandran, Martin Charles Golumbic, Rogers Mathew, and Deepak Rajendraprasad. Separation dimension of graphs and hypergraphs. *Algorithmica*, 75(1):187–204, 2016.

[BCM98] Andrei Z. Broder, Moses Charikar, and Michael Mitzenmacher. A derandomization using min-wise independent permutations. In Michael Luby, José

D. P. Rolim, and Maria Serna, editors, *Randomization and Approximation Techniques in Computer Science*, pages 15–24, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[BEI+12] Martin Brain, Esra Erdem, Katsumi Inoue, Johannes Oetsch, Jörg Pührer, Hans Tompits, and Cemal Yilmaz. Event-sequence testing using answer-set programming. *International Journal on Advances in Software*, 5(3&4), 2012.

[BJL99] Thomas Beth, Deiter Jungnickel, and Hanfried Lenz. *Design Theory*, volume 1 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2 edition, 1999.

[BM14] Jonathan E. Beagley and Walter Morris. Chromatic numbers of copoint graphs of convex geometries. *Discrete Mathematics*, 331:151–157, 2014.

[Bro97] Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE, 1997.

[BTI12] Mutsunori Banbara, Naoyuki Tamura, and Katsumi Inoue. Generating event-sequence test cases by answer set programming with the incidence matrix. In *Technical Communications of the 28th International Conference on Logic Programming (ICLP'12)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2012.

[CCHZ13] Yeow Meng Chee, Charles J. Colbourn, Daniel Horsley, and Junling Zhou. Sequence covering arrays. *SIAM Journal on Discrete Mathematics*, 27(4):1844–1861, 2013.

[CFT14] Andrea Calvagna, Andrea Fornaia, and Emiliano Tramontana. Combinatorial interaction testing of a Java Card static verifier. In *2014 IEEE Seventh International Conference on Software Testing, Verification and Validation Workshops*, pages 84–87. IEEE, 2014.

[CM10] Ondrej Chum and Jiri Matas. Large-scale discovery of spatially related images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):371–377, 2010.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[DM41] Ben Dushnik and Edwin W. Miller. Partially ordered sets. *American Journal of Mathematics*, 63:600–610, 1941.

[Dus50] Ben Dushnik. Concerning a certain set of arrangements. *Proceedings of the American Mathematical Society*, 1(6):788–796, 1950.

[EKR61] Paul Erdős, Chao Ko, and Richard Rado. Intersection theorems for systems of finite sets. *The Quarterly Journal of Mathematics*, 12(1):313–320, 01 1961.

[ES35] Paul Erdős and George Szekeres. A combinatorial problem in geometry. *Compositio mathematica*, 2:463–470, 1935.

[ES88] Paul H. Edelman and Michael E. Saks. Combinatorial representation and convex dimension of convex geometries. *Order*, 5(1):23–32, 1988.

[FK84] Michael L. Fredman and János Komlós. On the size of separating systems and families of perfect hash functions. *SIAM Journal on Algebraic Discrete Methods*, 5(1):61–68, 1984.

[Für96] Zoltán Füredi. Scrambling permutations and entropy of hypergraphs. *Random Structures & Algorithms*, 8(2):97–104, 1996.

[Für00] Zoltán Füredi. On the Prague dimension of Kneser graphs. In *Numbers, Information and Complexity*, pages 125–128. Springer, 2000.

[GG94] Mike Grannell and Terry Griggs. An introduction to Steiner systems. *Mathematical Spectrum*, 26(3):74–80, 1994.

[GHLS93] Niall Graham, Frank Harary, Marilynn Livingston, and Quentin F. Stout. Subcube fault-tolerance in hypercubes. *Information and Computation*, 102(2):280–314, 1993.

[Gin67] Boris D. Ginzburg. A certain number-theoretic function which has an application in coding theory. *Problemy Kibernetiki*, 19:249–252, 1967.

[GW22] Aidan R. Gentle and Ian M. Wanless. On perfect sequence covering arrays, February 4, 2022. preprint on arXiv:2202.01960.

[Har05] Alan Hartman. Software and hardware testing using combinatorial covering suites. In *Graph theory, combinatorics and algorithms*, pages 237–266. Springer, 2005.

[Hen06] Monika Henzinger. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, 2006.

[HLS10] Bing Han, Jimmy Leblet, and Gwendal Simon. Hard multidimensional multiple choice knapsack problems, an empirical study. *Computers & operations research*, 37(1):172–181, 2010.

[HM99] Serkan Hoşten and Walter D. Morris, Jr. The order dimension of the complete graph. *Discrete Math.*, 201(1-3):133–139, 1999.

[HWKK20] Linghuan Hu, W. Eric Wong, D. Richard Kuhn, and Raghu N. Kacker. How does combinatorial testing perform in the real world: an empirical study. *Empirical Software Engineering*, 25(4):2661–2693, 2020.

[Ind01] Piotr Indyk. A small approximately min-wise independent family of hash functions. *Journal of Algorithms*, 38(1):84–90, 2001.

[Ish95] Yoshiyasu Ishigami. Containment problems in high-dimensional spaces. *Graphs and Combinatorics*, 11(4):327–335, 1995.

[ITT00] Toshiya Itoh, Yoshinori Takei, and Jun Tarui. On permutations with limited independence. In *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 137–146, 2000.

[ITT03] Toshiya Itoh, Yoshinori Takei, and Jun Tarui. On the sample size of k-restricted min-wise independent permutations and other k-wise distributions. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 710–719, 2003.

[Iur22] Enrico Iurlano. Growth of the perfect sequence covering array number, 2022. submitted.

[Kat66] Gyula Katona. On separating systems of a finite set. *Journal of Combinatorial Theory*, 1(2):174–194, 1966.

[KHL+12] D. Richard Kuhn, James M. Higdon, James F. Lawrence, Raghu N. Kacker, and Yu Lei. Combinatorial methods for event sequence testing. In *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*, pages 601–609. IEEE, 2012.

[KKL12] David Kuhn, Raghu Kacker, and Yu Lei. Combinatorial testing, 2012. Encyclopedia of Software Engineering, [online] `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=910001` (last access: 2022-08-17).

[Kör73] János Körner. Coding of an information source having ambiguous alphabet and the entropy of graphs. In *6th Prague conference on information theory*, pages 411–425, 1973.

[KS73] Daniel J. Kleitman and Joel Spencer. Families of k-independent sets. *Discrete mathematics*, 6(3):255–262, 1973.

[KSS81] Daniel J. Kleitman, James B. Shearer, and Dean Sturtevant. Intersections of k-element sets. *Combinatorica*, 1(4):381–384, 1981.

[Lev91] Vladimir I. Levenshteĭn. Perfect codes in the metric of deletions and insertions. *Diskretnaya Matematika*, 3(1):3–20, 1991. Translation in Discrete Mathematics and Applications, 2 (1992), no. 3, 241–258.

[LK11] Ping Li and Arnd Christian König. Theory and applications of b-bit minwise hashing. *Communications of the ACM*, 54(8):101–109, 2011.

[LKL+11] Jim Lawrence, Raghu N. Kacker, Yu Lei, D. Richard Kuhn, and Michael Forbes. A survey of binary covering arrays. *The electronic journal of combinatorics*, pages P84–P84, 2011.

[Mar48] Edward Marczewski. Indépendance d'ensembles et prolongement de mesures (résultats et problèmes). In *Colloquium Mathematicum*, volume 2, pages 122–132, 1948.

[Mat97] Rudolf Mathon. Searching for spreads and packings. *London Mathematical Society Lecture Note Series*, pages 161–176, 1997.

[MGB+16] Nariman Mirzaei, Joshua Garcia, Hamid Bagheri, Alireza Sadeghi, and Sam Malek. Reducing combinatorics in GUI testing of Android applications. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, pages 559–570. IEEE, 2016.

[MKR12] Mohamed Z. Mohd Hazli, Zamli Kamal Z., and Othman Rozmie R. Sequence-based interaction testing implementation using bees algorithm. In *2012 IEEE Symposium on Computers & Informatics (ISCI)*, pages 81–85, 2012.

[MS03] Jiří Matoušek and Miloš Stojaković. On restricted min-wise independence of permutations. *Random Structures & Algorithms*, 23(4):397–408, 2003.

[Mul94] Ketan Mulmuley. *Computational Geometry. An introduction through randomized algorithms*. Prentice hall, 1994.

[MvT99] Rudolf Mathon and Tran van Trung. Directed t-packings and directed t-Steiner systems. *Designs, Codes and Cryptography*, 18(1):187–198, 1999.

[Na21] Jingzhou Na. Perfect sequence covering arrays. Master's thesis, Simon Fraser University, Canada, 2021.

[Nec65] Edward I. Nechiporuk. Complexity of gating circuits realized by Boolean matrices with undetermined elements. In *Doklady Akademii Nauk*, volume 163, pages 40–42. Russian Academy of Sciences, 1965.

[NJL22] Jingzhou Na, Jonathan Jedwab, and Shuxing Li. A group-based structure for perfect sequence covering arrays, February 4, 2022. preprint on arXiv:2202.01948.

[NZAA18] Abdullah B. Nasser, Kamal Z. Zamli, AbdulRahman A. Alsewari, and Bestoun S. Ahmed. An elitist-flower pollination-based strategy for constructing sequence and sequence-less t-way test suite. *International Journal of Bio-Inspired Computation*, 12(2):115–127, 2018.

[Rad01] Jaikumar Radhakrishnan. Entropy and counting. in *IIT Kharagpur Golden Jubilee Volume on Computational Mathematics, Modelling and Algorithms*. Ed. by J. C. Mishra (Narosa Publishers, New Delhi), 2001.

[Rad03] Jaikumar Radhakrishnan. A note on scrambling permutations. *Random Structures & Algorithms*, 22(4):435–439, 2003.

[Raj18] Deepak Rajendraprasad. Track number of line graphs. *Journal of Combinatorics*, 9(4):747–754, 2018.

[Rou87] Gilbert Roux. *k-Propriétés dans des tableaux de n colonnes : cas particulier de la k-surjectivité et de la k-permutivité*. PhD thesis, Université de Paris, 1987.

[RSK+20] Mostafijur Rahman, Dalia Sultana, Sabira Khatun, Mohd Falfazli Mat Jusof, Syamimi Mardiah Shaharum, Nurhafizah Abu Talip Yusof, Khandker M. Qaiduzzaman, Md Hasibul Hasan, Md Mushfiqur Rahman, Md Anwar Hossen, et al. *t*-way strategy for sequence input interaction test case generation adopting fish swarm algorithm. In *InECCE2019*, pages 87–99. Springer, 2020.

[SI22a] Neil J. A. Sloane and The OEIS Foundation Inc. Entry A000315 in The On-Line Encyclopedia of Integer Sequences [online], 2022. `https://oeis.org/A000315` (last access: 2022-06-20).

[SI22b] Neil J. A. Sloane and The OEIS Foundation Inc. Entry A053633 in The On-Line Encyclopedia of Integer Sequences [online], 2022. `https://oeis.org/search?q=A53633` (last access: 2022-04-26).

[Slo93] Neil J. A. Sloane. Covering arrays and intersecting codes. *Journal of combinatorial designs*, 1(1):51–63, 1993.

[Slo08] Neil J. A. Sloane. *On single-deletion-correcting codes*, pages 273–292. De Gruyter, 2008.

[Spe71] Joel Spencer. Minimal scrambling sets of simple orders. *Acta Mathematica Hungarica*, 22(3–4):349–353, 1971.

[Ste53] Jacob Steiner. Combinatorische Aufgaben. *Journal für die reine und angewandte Mathematik*, 45:181–182, 1853.

[SW20] Alex Scott and David R. Wood. Better bounds for poset dimension and boxicity. *Trans. Amer. Math. Soc.*, 373(3):2157–2172, 2020.

[Tar08] Jun Tarui. On the minimum number of completely 3-scrambling permutations. *Discrete mathematics*, 308(8):1350–1354, 2008.

[TIT03] Jun Tarui, Toshiya Itoh, and Yoshinori Takei. A nearly linear size 4-min-wise independent permutation family by finite geometries. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 396–408. Springer, 2003.

[Tro78] William T. Trotter. Some combinatorial problems for permutations. *Congressus Numerantium*, 19(1978):619–632, 1978.

[vLW01] Jacobus H. van Lint and Richard M. Wilson. *A Course in Combinatorics*. A Course in Combinatorics. Cambridge University Press, 2001.

[VT65] Rom R. Varshamov and Grigory M. Tenengolts. A code which corrects single asymmetric errors. *Avtomat. i Telemeh.*, 26:288–292, 1965.

[WLS+09] Wenhua Wang, Yu Lei, Sreedevi Sampath, Raghu Kacker, Rick Kuhn, and James Lawrence. A combinatorial approach to building navigation graphs for dynamic web applications. In *2009 IEEE International Conference on Software Maintenance*, pages 211–220. IEEE, 2009.

[Yan82] Mihalis Yannakakis. The complexity of the partial order dimension problem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):351–358, 1982.

[YCM07] Xun Yuan, Myra Cohen, and Atif M. Memon. Covering array sampling of input event sequences for automated GUI testing. In *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*, pages 405–408, 2007.

[Yus20] Raphael Yuster. Perfect sequence covering arrays. *Designs, Codes and Cryptography*, 88(3):585–593, 2020.

[ZMA16] Juan Zamora, Marcelo Mendoza, and Héctor Allende. Hashing-based clustering in high dimensional data. *Expert Systems with Applications*, 62:202–211, 2016.