Spanoids – An Abstraction of Spanning Structures, and a Barrier for LCCs

Zeev Dvir[1]
Dept of Computer Science and Dept of Mathematics, Princeton University, Princeton, NJ, USA
zeev.dvir@gmail.com

Sivakanth Gopi[2]
Microsoft Research, Redmond, WA, USA
sigopi@microsoft.com

Yuzhou Gu[3]
MIT, Cambridge, MA, USA
yuzhougu@mit.edu

Avi Wigderson[4]
Institute of Advanced Study, Princeton, NJ, USA
avi@math.ias.edu

Abstract

We introduce a simple logical inference structure we call a *spanoid* (generalizing the notion of a matroid), which captures well-studied problems in several areas. These include combinatorial geometry (point-line incidences), algebra (arrangements of hypersurfaces and ideals), statistical physics (bootstrap percolation), network theory (gossip / infection processes) and coding theory. We initiate a thorough investigation of spanoids, from computational and structural viewpoints, focusing on parameters relevant to the applications areas above and, in particular, to questions regarding Locally Correctable Codes (LCCs).

One central parameter we study is the *rank* of a spanoid, extending the rank of a matroid and related to the dimension of codes. This leads to one main application of our work, establishing the first known barrier to improving the nearly 20-year old bound of Katz-Trevisan (KT) on the dimension of LCCs. On the one hand, we prove that the KT bound (and its more recent refinements) holds for the much more general setting of spanoid rank. On the other hand we show that there exist (random) spanoids whose rank matches these bounds. Thus, to significantly improve the known bounds one must step out of the spanoid framework.

Another parameter we explore is the *functional rank* of a spanoid, which captures the possibility of turning a given spanoid into an actual code. The question of the relationship between rank and functional rank is one of the main questions we raise as it may reveal new avenues for constructing new LCCs (perhaps even matching the KT bound). As a first step, we develop an entropy relaxation of functional rank to create a small constant gap and amplify it by tensoring to construct a spanoid whose functional rank is smaller than rank by a polynomial factor. This is evidence that the entropy method we develop can prove polynomially better bounds than KT-type methods on the dimension of LCCs.

To facilitate the above results we also develop some basic structural results on spanoids including an equivalent formulation of spanoids as set systems and properties of spanoid products. We feel that given these initial findings and their motivations, the abstract study of spanoids merits further investigation. We leave plenty of concrete open problems and directions.

[1] Research supported by NSF CAREER award DMS-1451191 and NSF grant CCF-1523816.
[2] Research supported by NSF CAREER award DMS-1451191 and NSF grant CCF-1523816.
[3] Research supported by Jacobs Family Presidential Fellowship.
[4] Research was partially supported by NSF grant CCF-1412958.

10th Innovations in Theoretical Computer Science (ITCS 2019).
Editor: Avrim Blum; Article No. 32; pp. 32:1–32:20

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2012 ACM Subject Classification Theory of computation → Error-correcting codes

Keywords and phrases Locally correctable codes, spanoids, entropy, bootstrap percolation, gossip spreading, matroid, union-closed family

Digital Object Identifier 10.4230/LIPIcs.ITCS.2019.32

Related Version Full version at `https://arxiv.org/abs/1809.10372`.

Acknowledgements The second author would like to thank Sumegha Garg for helpful discussions. The third author would like to thank Yury Polyanskiy for helpful discussions.

1 Introduction

This (somewhat long) introduction will be organized as follows. We begin by discussing Locally Correctable Codes (LCCs) and the main challenges they present as this was the primary motivation for this work. We proceed to define spanoids as an abstraction of LCCs, and state some results about their rank which hopefully illuminate the difficulties with LCCs in a new light. We continue by describing other natural settings in which the spanoid structure arises in the hope of motivating the questions raised in the context of LCCs and demonstrating their potential to contribute to research in other areas. We then turn to the investigation of functional rank of spanoids, which aims to convert them to actual LCCs. We conclude with describing some of the structural results about spanoids obtained here.

1.1 Locally Correctable Codes

The introduction of *locality* to coding theory has created a large body of research with wide-ranging applications and connections, from probabilistically checkable proofs, private information retrieval, program testing, fault-tolerant storage systems, and many others in computer science and mathematics. We will not survey these, and the reader may consult the surveys [27, 10]. Despite much progress, many basic questions regarding local testing, decoding and correcting of codes remain open. Here we focus on the efficiency of locally *correctable* codes, that we now define. Note that the related, locally *decodable* codes (LDCs), will not be discussed in this paper, as our framework is not relevant to them (LCCs can be converted to LDCs with a small loss in parameters).

▶ **Definition 1** (q-LCCs)**.** A *code* $C \subseteq \Sigma^n$ is called a q-query locally correctable code with error-tolerance $\delta > 0$, if for every $i \in [n]$ there is a family (called a *q-matching*) M_i, of at least δn *disjoint* q-subsets of $[n]$, with the following *decodability* property. For every codeword $c \in C$, and for every $i \in [n]$, the value of c_i is *determined*[5] by the values of c in coordinates S, for every q-subset S in M_i.[6]

[5] Through some function that does not depend on the codeword c.

[6] Our definition is a 'zero-error' version of the standard definition. By 'zero-error' we mean that for any codeword c, the value of c_i can be determined correctly (without error) from the coordinates of c at any q-subset S in the matching M_i. A more general definition would say that c_i can be computed from $c|_S$ with high probability, or even just slightly better than a random guess. Our definition is equivalent to the more general definition for linear codes, which comprise all of the interesting examples. We still allow 'global' error in the sense that a large (constant) fraction of the coordinates can be corrupted (this global error tolerance is captured by the parameter δ).

Intuitively, given a vector $c' \in \Sigma^n$ which results from corrupting less than (say) $\epsilon\delta n$ coordinates of a codeword $c \in C$, recovering c_i for any given $i \in [n]$ is simple. Picking a random q-subset from M_i and decoding c_i according to it will succeed with probability at least $1 - \epsilon$, as only an ϵ-fraction of these q-subsets can be corrupted.

We focus in this paper on the most well-studied and well-motivated regime where both the "query-complexity" q and the error-tolerance δ are constants. It is not hard to see that there are no LCCs with $q = 1$ (unless the dimension is constant) and so we will start with the first interesting case of $q = 2$. A canonical example of a 2-query LCC, which will serve us several times below, is the Hadamard code. Here $\Sigma = \mathbb{F}_2$. Let k be any integer and set $n = 2^k - 1$. Let A be the $k \times n$ matrix whose columns are all non-zero k-bit vectors. The Hadamard code $C_H \in \mathbb{F}_2^n$ is *generated* by A, namely H consists of all linear combinations of rows of A. Since every column of A can be written as a sum of (namely, *spanned* by) pairs of other columns in $(n-1)/2$ different ways, the matching M_i suggest themselves, and so is the *linear* correcting procedure: add the values in coordinates of the random pair S from M_i to determine the ith coordinate.

A central parameter of codes is their *rate*, capturing the redundancy between the *dimension*, namely the number of information bits encoded (here k), and the length of the codeword (here n). As in this paper this k will be a tiny function of n, we will focus on the *dimension* itself. Note that in the example above, as in every *linear* code, this dimension is also the *rank* of the generating matrix. In general codes, dimension may be fractional, and is defined as follows. All logarithms are in base 2 unless otherwise noted.

▶ **Definition 2** (Dimension and rate of a code)**.** For a general, possibly non-linear code $C \subseteq \Sigma^n$, we define the dimension of C to be $\mathsf{dim}(C) = \log |C| / \log |\Sigma|$. Note that this coincides with the linear algebraic definition of dimension when C is a subspace. We refer to the ratio $\mathsf{dim}(C)/n$ as the 'rate' of the code.

Note that while the Hadamard code (C_H) has fantastic local correction (only 2 queries), its dimension is only $k \sim \log n$, which is pathetic from a coding theory perspective. However, no better 2-query LCC can exist, regardless of the alphabet.

▶ **Theorem 3** (2-LCCs)**.** *For all large enough n and over any alphabet:*

- *There exists a 2-query LCC of dimension $\Omega(\log n)$ and constant δ (Folklore: Hadamard code).*
- *Every 2-query LCC must have dimension at most $O(\log n)$ (for any constant δ) [6].*

While we know precisely the optimal dimension for 2 queries, for $q \geq 3$ the gap between known upper and lower bounds is huge. The best lower bounds (constructions) are polylogarithmic: they come from Reed-Muller codes (using polynomials over finite fields), and yield dimension $\Omega((\log n)^{q-1})$.

The best LCC upper bounds are only slightly sub-linear, giving $\mathsf{dim}(C) \leq \widetilde{O}(n^{1-\frac{1}{q-1}})$ (up to logarithmic factors). This bound, which we will refer to as the Katz-Trevisan (KT) bound, is actually a slight refinement/improvement over the bound originally appearing in [19] (which gave $n^{1-1/q}$). This improvement was implicit in several works (e.g. [9, 25]) and is explicitly stated in [18]. We should also note that, over constant-size alphabets, Kerenidis and De-Wolf proved an even stronger bound using quantum information theory [21]. This exponential gap between the upper and lower bounds, which we formally state below, has not been narrowed in over two decades.[7] Explaining this gap (in the hope of finding ways to close it) is one major motivation of this work.

[7] For LDCs better constructions than Reed-Muller codes are known, through the seminal works of [26, 12], but as mentioned we will not discuss them here. Still, the upper bounds for LDCs are the same as for

▶ **Theorem 4** (q-LCCs, $q \geq 3$)**.** *For every fixed $q \geq 3$ and all large enough n:*

- *There exists a q-query LCC of dimension $\Omega((\log n)^{q-1})$ (with constant δ and alphabet of size $q+1$) (Reed-Muller codes, see e.g. the survey [27]).*
- *Every q-query LCC must have dimension at most $\widetilde{O}(n^{1-\frac{1}{q-1}})$ (for any constant δ and any alphabet) [18].*

1.2 Spanoids

We shall now abstract the notion of inference used in LCCs. There, for a collection of pairs (S, i) with $S \subseteq [n]$ and $i \in [n]$, the values of codewords in coordinate positions S, determine the value of of some other coordinate i. We shall forget (for now) the underlying code altogether, and abstract this relation by the formal "inference" symbol $S \to i$, to be read "S *spans* i".

▶ **Definition 5** (Spanoid)**.** A *spanoid* $\mathcal{S}$ over $[n]$ is a family of pairs (S, i) with $S \subseteq [n]$ and $i \in [n]$. The pair (S, i) will sometimes be written as $S \to i$ and read as S spans i in the spanoid $\mathcal{S}$.

One natural way to view a spanoid is as a logical inference system, with the pairs indicating all inference rules. The elements of $[n]$ indicate some n formal statements, and an inference $S \to i$ of the spanoid means that if we know the truth of the statements in S, we can infer the truth of the ith statement. With this intuition, we shall adopt the convention that the inferences $i \to i$ are implicit in any spanoid, and that *monotonicity* holds: if $S \to i$ then also $S' \to i$ for every $S' \supseteq S$. These conventions will be formally stated below when we define general derivations, which sequentially combine these implicit rules and the stated rules (pairs) of the spanoid.

A key concept of spanoids is, naturally, the *span*. Given a subset $T \subset [n]$ (which we can think of as "axioms"), we can explore everything they can span by a sequence of applications of the inference rules of the spanoid $\mathcal{S}$.

▶ **Definition 6** (Derivation, Span)**.** A *derivation* in $\mathcal{S}$ of $i \in [n]$ from $T \subseteq [n]$, written $T \models_{\mathcal{S}} i$, is a sequence of sets $T = T_0, T_1, \ldots, T_r$ with $i \in T_r$ such that for each $j \in [r]$, $T_j = T_{j-1} \cup i_j$ for some $i_j \in [n]$ and there exists $S \subset T_{j-1}$ such that $(S, i_j) \in \mathcal{S}$ is one of the spanoid rules.

The *span* (or *closure*) of T, denoted $\mathsf{span}_{\mathcal{S}}(T)$, is the set of all i for which $T \models_{\mathcal{S}} i$. We shall remove the subscript $\mathcal{S}$ from these notations when no confusion about the underlying spanoid can arise, and write $T \models i$ and $\mathsf{span}(T)$ for short.

Despite being highly abstract, we will see that spanoids can lead to a rich family of questions and definitions. The first, and perhaps one of the most central definitions is that of the *rank* of a spanoid. We shall see other notions of spanoid rank later on (and will discuss the relation between them).

▶ **Definition 7** (Rank)**.** The *rank* of a spanoid $\mathcal{S}$, denoted $\mathsf{rank}(\mathcal{S})$, is the size of the *smallest* subset $T \subseteq [n]$ such that $\mathsf{span}(T) = [n]$. Note that by the definition of span we always have $\mathsf{rank}(\mathcal{S}) \leq n$.

We note that the "rank" of a logical inference system does appear (under different names) in *proof complexity*. It is the starting point for *expansion-based* lower bounds on a variety of proof systems, as introduced for Resolution proofs in [5], and used for many others e.g. in [1] and [2]). We shall return to this connection presently.

LCCs, and obtained by the same KT-type argument, so the results in this paper may serve to better understand the (smaller, but still quite large) gap between upper and lower bounds in LDCs as well.

We can now define the spanoid analog of q-LCCs as spanoids which only specify the correction structure (the matchings M_i) without requiring any codewords or alphabet.

▶ **Definition 8** (q-LCS, Locally correctable spanoid)**.** A spanoid $\mathcal{S}$ over $[n]$ is a q-LCS with error-tolerance δ if for every $i \in [n]$ there exists a family M_i of at least δn disjoint q-subsets of $[n]$ such that for each $S \in M_i$ we have $(S, i) \in \mathcal{S}$. Namely, each $i \in [n]$ is spanned (in $\mathcal{S}$) by at least δn disjoint subsets of q-elements.

One can now ask about the highest possible rank of a q-LCS. It is not hard to see that the existence of a q-LCC (over any alphabet) $C \subset \Sigma^n$ with dimension $\mathsf{dim}(C) = d$ automatically implies that there exists a q-LCS (namely, the one given by the same matchings used in C) with rank at least $\lceil d \rceil$. Indeed, otherwise there would be $r < d$ coordinates in $[n]$ that determine any codeword $c \in C$ and this would limit the number of codewords to Σ^r.

One of our main observations is that, remarkably, in locally correctable spanoids there is *no* gap between the upper and lower bounds: we know the precise answer up to logarithmic factors, and it matches the *upper bounds* for LCCs! Observe the analogies to the theorems in the previous subsection, for $q = 2$ and $q \geq 3$.

▶ **Theorem 9** (2-LCSs)**.** *For all large enough n:*

- *There exists a 2-LCS over $[n]$ with error-tolerance δ of rank $\Omega(\frac{1}{\delta}\log(\delta n))$.*
- *Every 2-LCS over $[n]$ with error-tolerance δ must have rank at most $O\left(\frac{1}{\delta}\log(n)\right)$.*[8]

Here, of course, the inference structure of the Hadamard code proves the first item. To get the required dependence on δ, one can take $\frac{1}{\delta}$ disjoint copies of such spanoids. The second item requires a new proof we discuss below, which generalizes (and implies) the one in Theorem 3. It is quite surprising that, even in this abstract setting, with no need for codewords or alphabet, one cannot do better than the Hadamard code!

We now state our results for $q \geq 3$.

▶ **Theorem 10** (q-LCSs with $q \geq 3$)**.** *For every fixed $q \geq 3$ and all large enough n:*

- *There exist a q-LCS of rank $\tilde{\Omega}(n^{1-\frac{1}{q-1}})$ (with constant δ).*
- *Every q-LCS over $[n]$ has rank at most $\tilde{O}\left(n^{1-\frac{1}{q-1}}\right)$ (for any constant δ).*

Both parts of this theorem demand discussion. The possibly surprising (and tight) lower bound follows from a simple probabilistic argument (indeed, one which is repeatedly used to prove expansion in the proof complexity references cited above), where the matchings M_i are simply chosen uniformly at random. It seems to reveal how significant a relaxation spanoids are of LCCs (where probabilistic arguments fail completely). However, the best known LCC upper bound (Theorem 4) does not rule out the possibility that, at least for large alphabets, the two (LCC's dimension and LCS's rank) have the same behavior! From a more pessimistic (and perhaps more realistic) perspective, our lower bound shows the limitations of any (upper bound) proof technique which, in effect, applies also for spanoids. These are proofs in which the LCC structure is used to show that a small subset spans all the others. We note that there are several LCC upper bounds which 'beat' the $n^{1-\frac{1}{q-1}}$ bound for certain very special cases by using additional structure not present in the corresponding abstract spanoid. One example is the bound of [21], which uses arguments from quantum information theory to roughly *halve* the number of queries, over binary (or small) alphabets. Another example is

[8] The results of [6] can be interpreted as an upper bound of $O\left(\mathsf{poly}(1/\delta)\log(n)\right)$ on the rank of 2-LCS with error-tolerance δ.

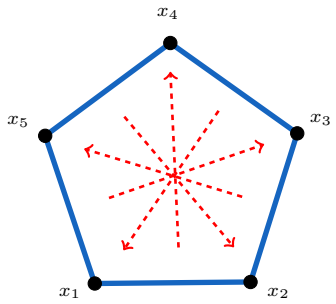

Figure 1 The pentagon spanoid Π_5 where each coordinate is spanned by the coordinates of the opposite edge.

the paper [11], which gives an improved upper bound on the dimension for linear 3-LCCs defined over the real numbers, using specific properties of the Reals such as distance and volume arguments.

Our proof of the upper bound, is again more general than for LCCs, and interesting in its own right. We use a simple technique which performs random restrictions and contractions of graphs and hypergraphs (and originates in [11]). It will be described in Section 3, after we have formulated an equivalent, set-theoretic formulation of spanoids in Section 2.2.

1.2.1 Functional rank: bridging the gap between LCCs and LCSs

We conclude this section of the introduction with an attempt to understand (and possibly bridge) the gap between LCCs and their spanoid abstraction. The idea is to start with an LCS of high rank (which we know is possible), and convert it to an LCC without losing too much in the parameters. More generally, for a given spanoid $\mathcal{S}$, we would like to investigate the code C with largest dimension (over any alphabet Σ) which would be consistent with the inferences of $\mathcal{S}$. This is captured in the notion of *functional rank* which we now define.

▶ **Definition 11** (Functional rank)**.** Let $\mathcal{S}$ be a spanoid over $[n]$. A code $C \subset \Sigma^n$ is *consistent* with $\mathcal{S}$ if for every inference (S, i) in $\mathcal{S}$, and for every codeword $c \in C$, its values of coordinates S determine its value in coordinate i (by some fixed function, $f_{S,i}$ not depending on c).[9]

Define the *functional rank* of $\mathcal{S}$, denoted $\mathsf{f\text{-}rank}(\mathcal{S})$, to be equal to the supremum of the dimension $\mathsf{dim}(C)$, over all possible finite alphabets Σ and codes $C \subset \Sigma^n$ which are consistent with $\mathcal{S}$.

Of course, the strategy of constructing LCCs in two stages as above can only work if we can bound the gap between $\mathsf{rank}(\mathcal{S})$ and $\mathsf{f\text{-}rank}(\mathcal{S})$. This question, of bounding this gap or proving it can be large, is perhaps the most interesting one we raise (and leave mostly open for now). For now, we are able to show an example in which the two are different. The example providing a gap is depicted in Figure 1, arranging the coordinates as the vertices of a pentagon, the pair of vertices of each edge span the vertex opposite to it. That is, $\{x_1, x_2\} \to x_4, \{x_2, x_3\} \to x_5$ etc.

▶ **Theorem 12** (Constant gap between rank and functional rank)**.** *The pentagon spanoid* Π_5 *depicted in Figure 1 has* $\mathsf{rank}(\Pi_5) = 3$ *but* $\mathsf{f\text{-}rank}(\Pi_5) = 2.5$.

[9] One can think of a code consistent with $\mathcal{S}$ also as a 'representation' of $\mathcal{S}$ in the spirit of matroid theory.

Seeing that $\mathsf{rank}(\Pi_5) = 3$ is easy by inspection. The lower bound of 2.5 on functional rank comes from a set-theoretic construction of consistent codes. This comes from a simple linear programming (LP) relaxation we develop for $\mathsf{rank}(\mathcal{S})$ called $\mathsf{LP}^{cover}(\mathcal{S})$, but surprisingly this LP captures the best set-theoretic construction of consistent codes. But even in this small example, the upper bound on functional rank is nontrivial to determine, as we allow all possible alphabets and consistent codes. Not surprisingly, Shannon entropy is the key to proving such a bound. We develop a linear programming relaxation, based on entropy whose optimum $\mathsf{LP}^{entropy}(\mathcal{S})$ *upper bounds* $\mathsf{f\text{-}rank}(\mathcal{S})$. In this example, it proves 2.5 to be the optimum. For the definitions of $\mathsf{LP}^{entropy}(\mathcal{S})$, $\mathsf{LP}^{cover}(\mathcal{S})$ and the proof of Theorem 12 see the full version. One natural way of amplifying gaps as in the example above, which may also be useful in creating codes of high functional rank, is the idea of *tensoring*. We develop different notions of tensoring *spanoids* inspired by tensoring of codes. In particular, we define a product of spanoids called the *semi-direct product* under which rank is multiplicative and $\mathsf{f\text{-}rank}$ is sub-multiplicative. By repeatedly applying this product to Π_5, we get a spanoid with polynomial gap between $\mathsf{f\text{-}rank}$ and rank. See the full version for details.

▶ **Theorem 13** (Polynomial gap between rank and function rank). *There exists a spanoid $\mathcal{S}$ on n elements with $\mathsf{rank}(\mathcal{S}) \geq n^c \mathsf{f\text{-}rank}(\mathcal{S})$ where $c = \log_5 3 - \log_5 2.5 \geq 0.113$.*

Summarizing, we have the following obvious inequalities between the measures we described so far for every spanoid $\mathcal{S}$. We feel that understanding the exact relationships better is worthy of further study

$$\mathsf{LP}^{cover}(\mathcal{S}) \leq \mathsf{f\text{-}rank}(\mathcal{S}) \leq \mathsf{LP}^{entropy}(\mathcal{S}) \leq \mathsf{rank}(\mathcal{S}). \tag{1}$$

1.3 Other motivations and incarnations of spanoids

We return to discuss other structures, combinatorial, geometric and algebraic, in which the same notions of span and inference naturally occur, leading to a set-theoretic one that elegantly captures spanoids precisely. These raise further issues, some of which we study in this paper and some are left for future work. These serves to illustrate the breadth of the spanoid framework.

1.3.1 Bootstrap percolation and gossip processes

The following general set-up occurs in statistical physics, network theory and probability theory. Fix an undirected graph $G([n], E)$. In a gossip or infection process, or equivalently bootstrap percolation, we are give a set of "rules" specifying, for every vertex $v \in [n]$, a family of subsets of its neighbors. The intended meaning of such a rule is that if *every* member of one such subset is "infected" at a certain time step, then the vertex v becomes infected in the next time step. Given a set of initial infected vertices, this defines a process in which infection spreads, and eventually stabilizes. A well studied special case is the (uniform) *r-bond percolation* [7], where the family for each vertex is all r-subsets of its neighbors. Many variants exist, e.g. one can have a similar process on the edges, rather than vertices of the graph. An important parameter of such a process is the following: what is the size of the smallest set of vertices which, if infected, will eventually infect all other vertices[10].

[10] This turns out to be crucial for understanding, at least for certain structured graphs like lattices studied by physicists, the threshold probability for percolation when initial infections are random.

A moment's thought will convince the reader that this structure is precisely a spanoid (where inferring sets are restricted by the graph structure). The infection process is precisely the inference process defining span in spanoids. Furthermore, the smallest size of an infecting set is *precisely* the rank of that spanoid! Much work has been invested to determine that rank even in very special cases, e.g. for the r-bond percolation above, in e.g. Boolean hypercubes, where it is known precisely. Interestingly, the paper [16] uses the so-called "polynomial method" to reprove that bound, which fits even deeper with our framework. In our language, their method determines the *functional rank* of this spanoid, and one direction is through constructing an explicit code that is consistent with the spanoid! The reader is encouraged to work out the details.

1.3.2 Independence systems and Matroids

An *independence system* over $[n]$ is a family $\mathcal{F}$ of subsets of $[n]$ which is downwards-closed (if a set is in $\mathcal{F}$, so are all its subsets). The members of $\mathcal{F}$ are called *independent*. While much of what we say below generalizes to all independence systems, we specify them for the important special systems called *matroids*.

A *matroid* is an independence system in which the independent sets satisfy the so-called "exchange axiom" (which we will not define here). Matroids abstract *linear* independence in subsets of a vector space over a field[11], and capture algorithmic problems in which optimization is possible through the greedy algorithm. Matroids thus come with natural notions of span and rank, extending the ones in the linear algebraic setting. The rank of a set is the size of the largest independent set it contains. The span of a set is the maximal superset of it of the same rank. A matroid can thus be naturally viewed as a spanoid, with the inference rules $F \to i$ for every independent $F \in \mathcal{F}$ and every i for which $F \cup \{i\}$ is *not* independent (such minimal dependent sets as $F \cup \{i\}$ are called *cycles*). It is easy to verify that the notions of span and rank of the matroid and the spanoid it defines coincide. This also raises the natural question of bounding the gap between $\mathsf{f\text{-}rank}$ and rank for the special case of spanoids arising from matroids.

Note that a spanoid resulting from a matroid this way is *symmetric*: by the exchange property of matroids, if $E \subset [n]$ is a cycle of $\mathcal{F}$, then for *every* $i \in E$ it contains the inference $E \setminus \{i\} \to i$. Symmetric spanoids are interesting, and we note that the pentagon example witnessing the gap between rank and functional rank is *not* symmetric, and we do not know such a gap for symmetric spanoids. We also don't know if symmetric spanoids can achieve the lower bound in Theorem 10.

1.3.3 Point-line incidences

Sylvester-Gallai theorem is a celebrated result in combinatorial geometry conjectured by Sylvester and proved independently by Melchior and Gallai. It states that for any set of n points in Euclidean space $\mathbb{R}^d$, if the line through any two points passes through a third point, then they must all be collinear (namely, they span a 1-dimensional affine space). Over the complex numbers, one can prove a similar theorem but with the conclusion that the points span a 2-dimensional affine space (and there are in fact two dimensional examples known) [20]. Over finite fields the conclusion is even weaker, saying that the span has dimension at most $O(\log(n))$ and this is tight as the example of all points in $\mathbb{F}_p^k$ with $n = p^k$ shows. It

[11] Matroids are in fact more general than linear independent sets of vectors over a field, for example the Vámos matroid on eight elements is not representable over any field.

is not a coincidence that this example reminds one of the Hadamard code described before as an example of a 2-query LCC. It is in fact true that there is a tight connection between configurations of points with many collinear triples and linear 2-query LCCs. This was first noticed in [4, 3] and was used to prove that 2-LCCs do not exist over the characteristic zero fields (for $q \geq 3$ these questions are wide open with even larger gaps than in the finite field case). LCCs with more than 2 queries naturally correspond to point configurations with many $(q-2)$-dimensional affine spaces containing at least q points.

Given the connections between Sylvester-Gallai type incidence structures and LCCs, and the insights offered by spanoids for studying LCCs, it is natural that the study of the spanoid structures can help us get new insights on incidence geometry problems. A *dual* way to view these incidences, which we shall presently generalize, is to consider each point $p_i : i \in [n]$ as representing a hyperplane F_i through the origin (in the appropriate vector space) vanishing on the linear function defined by p_i. A point p_i is spanned by a collection of points $\{p_j : j \in S\}$ iff $F_i \supset \cap_{j \in S} F_j$. Therefore the spanning structure of the points $p_1, p_2, \ldots, p_n$ is captured by the spanoid where we would add the inference $S \to i$ iff F_i *contains* the common intersection of all $F_j : j \in S$.

1.3.4 Systems of polynomial equations

Given the above example, there is no reason to stop at the linear setting. Instead of lines we can consider n (multivariate) polynomials f_i over a field, and again consider F_i to be the zero set of f_i. The spanoid above, having an inference $S \to i$ whenever the set F_i *contains* the common intersection of all $F_j : j \in S$, is capturing another natural algebraic notion. Namely, it says that the polynomial f_i vanishes on the all the common roots of the polynomials $\{f_j : j \in S\}$. By the celebrated Hilbert's Nullstellensatz theorem, over algebraically closed fields, this implies that f_i belongs to the radical of the ideal generated by the f_j's. Here the rank function is far from being that of a matroid; the complex spanoid which arises (and in general is far from understood) plays a role in arithmetic complexity (a beautiful example is the recent [23] dealing with degree-2 polynomials).

1.3.5 Intersecting set systems

Let us remove all restrictions from the origin or nature of the n sets F_i discussed in the previous discussion. Assume we are given any such family $\mathcal{F}$ of sets (from an arbitrary universe, say U). As above, a natural spanoid $\mathcal{S}_{\mathcal{F}}$ will have the inference $S \to i$ whenever the set F_i *contains* the common intersection of all $F_j : j \in S$. Such situations (and hence, spanoids) arise in many questions of extremal set theory, for example the study of (weak) sunflowers, or families with certain forbidden intersection (or union) patterns, e.g. [14, 13, 15].

What is interesting in this more general framework, where the initial family of sets $\mathcal{F}$ is arbitrary, is that it becomes *equivalent* to spanoids! In other words, every spanoid $\mathcal{S}$ arises as $\mathcal{S}_{\mathcal{F}}$ of *some* family of sets $\mathcal{F}$. This possibly surprising fact is not much more than an observation, but it turns out to be an extremely useful formulation for proving some of the results in this paper. Let us state it formally (it will be proved in Section 2.2).

▶ **Theorem 14** (Spanoids and intersecting sets). *Let $\mathcal{S}$ be any spanoid on $[n]$. Then, there exists a universe U and a family $\mathcal{F}$ of n sets of U, $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$, such that $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$.*

It is convenient to assume that the sets in $\mathcal{F}$ have no element in common to all[12].

The notions of rank and span are extremely simple in this set-theoretic setting, and do not require the sequential "derivation" and the implicit ordering which we require to define these in spanoids. For a family $\mathcal{F}$ of n sets and a subset $S \subset [n]$, let us denote by $\cap S$ the subset of U which is the intersection of all $\{F_j : j \in S\}$. Then the rank of S is the size of the smallest subset $S' \subseteq S$ for which $\cap S' = \cap S$. Similarly, the span of S is the largest superset $S'' \supseteq S$ for which $\cap S'' = \cap S$.

These static definitions of rank and span make many things transparent. For example, the expected fact that testing if the rank of a spanoid (namely the rank of the set $[n]$) is at most some given integer k is $NP-complete$ (Claim 22). Complementing the sets F_i in $\mathcal{F}$, and replacing intersection with union, this is precisely the Set Cover problem. This connection also underlies the cover-based linear program discussed earlier, as well as proofs of the main quantitative results Theorem 9 and Theorem 10.

1.3.6 Union-closed families

Spanoids over $[n]$ are equivalent to union-closed families of subsets of $[n]$ i.e. a family of subsets of $[n]$ such that the union of any two members is again in the family. A closed set of a spanoid $\mathcal{S}$ is a subset $A \subset [n]$ such that $\mathsf{span}(A) = A$. The family of all closed sets of a spanoid $\mathcal{S}$ is denoted by $\mathcal{C}_\mathcal{S}$ which is an intersection-closed family. Thus the family of all open sets which are complements of closed sets is a union-closed family and is denoted by $\mathcal{O}_\mathcal{S}$. One can construct all the derivations of the spanoid given its family of open or closed sets. Conversely, given any union-closed family of subsets of $[n]$, one can define a spanoid whose open sets are precisely the given family. Thus spanoids on $[n]$ are equivalent to union-closed families of subsets of $[n]$. The rank of a spanoid has a very simple interpretation in terms of its open sets, $\mathsf{rank}(\mathcal{S})$ is equal to the size of the smallest hitting set for its family of open sets $\mathcal{O}_\mathcal{S}$. Moreover, $\mathsf{rank}(\mathcal{S})$ is at most $\log|\mathcal{O}_S|$. These connections are discussed in Section 2.1.

Union-closed families are interesting combinatorial objects with a rich structure. The widely open Frankl's union-closed conjecture states that in every union-closed family of N sets, there exists an element which is contained in at least $N/2$ sets. Though this was proved for various special classes (see survey [8]), the best general bound is $\Omega(N/\log N)$ due to [22, 24]. When seen in the framework of spanoids, this follows immediately from Claim 19 which says that there is a $\log N$ sized hitting set for every union-closed family of N sets. Thus there is an element which should hit at least $N/\log_2(N)$ sets. We hope that viewing union-closed families as spanoids could be of use in understanding them.

1.4 Organization

In Section 2, we will present two alternative ways to represent spanoids that will be very useful. In Section 3, we show upper bounds on the rank of q-LCSs for $q \geq 2$ thus proving the upper bounds in Theorems 27 and 29. In Section 4, we construct q-LCSs thus proving the lower bound in Theorem 29.

[12] Indeed, otherwise we can remove the common intersection (if non-empty) of *all* members of $\mathcal{F}$ from each of them, as it does not change the underlying spanoid.

2 Preliminaries on spanoids

We will now describe two equivalent ways to define spanoids which will turn out to be very useful.

2.1 Spanoids as union-closed or intersection-closed families

In this subsection, we will define an equivalent and more canonical way of describing spanoids in terms of intersection closed or union closed families. We have defined spanoids in Definition 5 by specifying the initial set of derivation rules such as $A \to i$. But two different initial set of rules can lead to the same set of derivations and we should consider two spanoids to be equivalent if they lead to the same set of derivations. We will present an alternative way to describe spanoids which makes them equivalent to union-closed families of sets or alternatively intersection-closed families of sets. Moreover this new representation is a more canonical way to represent spanoids since it will be based only on the set of derivations. For this, the main new notions we need to define are that of a 'closed set' and an 'open set'.

▶ **Definition 15.** (Closed and open sets) Let $\mathcal{S}$ be an spanoid on $[n]$. A *closed set*[13] is a subset $B \subset [n]$ for which $\mathsf{span}(B) = B$. A subset $B \subset [n]$ is called an *open set* if its complement is a closed set. The family of all closed sets of $\mathcal{S}$ is denoted by $\mathcal{C}_\mathcal{S}$ and the family of all open sets of $\mathcal{S}$ by $\mathcal{O}_\mathcal{S}$ (when it is clear from the context, we will drop the subscript $\mathcal{S}$).

▶ **Claim 16.** *In any spanoid $\mathcal{S}$ on $[n]$,*

1. *the intersection of any number of closed sets is a closed set i.e. $\mathcal{C}_\mathcal{S}$ is an intersection-closed family,*
2. *the union of any number of open sets is an open set i.e. $\mathcal{O}_\mathcal{S}$ is a union-closed family and*
3. *for any set $A \subset [n]$, $\mathsf{span}(A)$ is equal to the intersection of all closed sets containing A i.e. $\mathsf{span}(A) = \bigcap_{B \supset A, B \in \mathcal{C}_\mathcal{S}} B$.*

Proof.

(1) Let $F = F_1 \cap F_2$ be the intersection of two closed sets. Suppose in contradiction that F spans some element $x \in [n] \setminus F$ then, by monotonicity, both F_1 and F_2 have to span x. Hence, $x \in \mathsf{span}(F_1) \cap \mathsf{span}(F_2) = F_1 \cap F_2 = F$ in contradiction.

(2) This just follows from (1) by taking complements.

(3) Let $F(A)$ be the intersection of all closed sets containing A. Since $\mathsf{span}(A)$ is a closed set we clearly have $F(A) \subset \mathsf{span}(A)$. To see the other direction, suppose $x \in \mathsf{span}(A)$ and let F be any closed set containing A. Then, by monotonicity, F must also span x and so we must have $x \in F$. ◀

▶ **Claim 17.** *A spanoid is uniquely determined by the set of all its closed (open) sets which is an intersection-closed (union-closed) family of subsets of $[n]$. Conversely, every intersection-closed (union-closed) family of subsets of $[n]$ defines a spanoid whose closed (open) sets are the given family.*

Proof. Given a spanoid $\mathcal{S}$ on $[n]$, by Claim 16, we can define $\mathsf{span}(A)$ in $\mathcal{S}$ using just the closed sets as:

$$\mathsf{span}(A) = \bigcap_{B \supset A, B \in \mathcal{C}_\mathcal{S}} B.$$

[13] Closed sets are analogous to 'flats' or 'subspaces' in matroids.

And $A \models i$ in $\mathcal{S}$ iff $i \in \mathsf{span}(A)$. Thus given the set of all closed sets, we can reconstruct all the derivations of the spanoid.

For the converse, suppose we are given an intersection-closed family of subsets of $[n]$, say $\mathcal{C}$. We can define $\mathsf{span}_{\mathcal{C}}(A) = \bigcap_{B \supset A, B \in \mathcal{C}} B$ and define a spanoid $\mathcal{S}_{\mathcal{C}}$ where $A \models i$ iff $i \in \mathsf{span}_{\mathcal{C}}(A)$. It is easy to see that the closed sets of this spanoid $\mathcal{S}_{\mathcal{C}}$ is exactly $\mathcal{C}$. ◀

Thus an equivalent way to define a spanoid is to define all its closed (open) sets which is some intersection (union) closed family. The following claim shows that the rank of a spanoid has a very natural interpretation in terms of the open sets.

▶ **Claim 18.** *The rank of a spanoid $\mathcal{S}$ is the size of the smallest hitting set for the collection $\mathcal{O}_{\mathcal{S}}$ i.e. a set which intersects every open set in $\mathcal{O}_{\mathcal{S}}$ non-trivially.*

Proof. Observe that a subset $A \subset [n]$ spans $[n]$ iff it is a hitting set for all the open sets in $\mathcal{O}_{\mathcal{S}}$. This is because if A doesn't hit some open set B, then A lies in the complement of B i.e. $A \subset \bar{B}$. Since $\bar{B}$ is closed, $\mathsf{span}(A) \subset \bar{B} \neq [n]$. Therefore $\mathsf{rank}(\mathcal{S})$ is the size of the smallest hitting set for $\mathcal{O}_{\mathcal{S}}$. ◀

This interpretation of the rank is used to give a linear programming relaxation LP^{cover} which lower bounds the rank. We can also upper bound the rank of a spanoid in terms of the number of closed or open sets as the following claim shows.

▶ **Claim 19.** *Let $\mathcal{S}$ be a spanoid, then $\mathsf{rank}(\mathcal{S}) \leq \log_2(|\mathcal{C}_{\mathcal{S}}|) = \log_2(|\mathcal{O}_{\mathcal{S}}|)$.*

Proof. Let $r = \mathsf{rank}(\mathcal{S})$ and $R \subset [n]$ be a set of size $|R| = r$ spanning $[n]$. Since the rank of $\mathcal{S}$ is r we know that R is independent (not spanned by any proper subset). For each of the 2^r subsets $S \in 2^R$ we consider the closed set $F_S = \mathsf{span}(S)$. We claim that all of these are distinct. Suppose in contradiction that there were two distinct sets $S \neq T \in 2^R$ with $\mathsf{span}(S) = \mathsf{span}(T)$. W.l.o.g suppose there is an element $x \in T \setminus S$. Then $x \in \mathsf{span}(S)$ and so we get that $R \setminus \{x\}$ spans R (by monotonicity) and so spans the entire spanoid in contradiction. Thus $|\mathcal{C}_{\mathcal{S}}| \geq 2^r$. ◀

2.2 Spanoids as set systems

In this subsection, we will show yet another way of representing spanoids by families of sets. This representation (which is equivalent to spanoids) will be easier to work with and, in fact, we will later work almost exclusively with it instead of with the definition given in the introduction. Recall the notation introduced at the end of the introduction that, for sets $S_1, \ldots, S_n$ and for a subset $A \subset [n]$ we let $\cap A = \cap_{i \in A} S_i$.

▶ **Definition 20** (Intersection Dimension of a set system)**.** The *intersection-dimension* of a family of sets $S_1, \ldots, S_n$, denoted $\mathsf{idim}(S_1, \ldots, S_n)$ is the smallest integer d such that there exist a set $A \subset [n]$ of size d such that $\cap A = \cap[n]$.

▶ **Lemma 21** (Set-Representation of spanoids)**.** *Let $\mathcal{S}$ be a spanoid on $[n]$ with $\mathsf{rank}(\mathcal{S}) = r$. Then there exists a family of sets $S_1, \ldots, S_n$ such that $A \models i$ in $\mathcal{S}$ iff $\cap A \subset S_i$. In this case we say that the set family $(S_1, S_2, \ldots, S_n)$ is a* set-representation *of $\mathcal{S}$ and this implies in particular that $\mathsf{idim}(S_1, \ldots, S_n) = \mathsf{rank}(\mathcal{S})$.*

Proof. For $i \in [n]$ we define $S_i \subset \mathcal{C}_{\mathcal{S}}$ to be the subfamily of closed sets of $\mathcal{S}$ containing the element $i \in [n]$. For the first direction of the proof suppose that A spans x in the spanoid $\mathcal{S}$. Then, by Claim 16, x belongs to any closed set containing A and so $\cap_{i \in A} S_i \subset S_x$. For the other direction, suppose $\cap_{i \in A} S_i \subset S_x$ or that any closed set containing A must also contain x. Hence, x is in the intersection of all closed sets containing A and, by Claim 16 we have that $x \in \mathsf{span}(A)$. ◀

An alternative way to represent spanoids is by unions. $(T_1, T_2, \ldots, T_n)$ is called a *union set-representation* of the spanoid $\mathcal{S}$ when, $A \models i$ in $\mathcal{S}$ iff $T_i \subset \cup_{j \in A} T_j$. Note that if $(S_1, S_2, \ldots, S_n)$ is an (intersection) set-representation for $\mathcal{S}$ as in Lemma 21, then by taking complements, $(\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_n)$ is a union set-representation for $\mathcal{S}$ and vice versa. Thus these two notions of representing a spanoid by sets is equivalent.

▶ **Claim 22.** *Given a spanoid S and some positive integer k, deciding if the rank of the spanoid is at most k is NP-complete.*

Proof. Given the description of a spanoid and a subset of its elements, we can check in polynomial time whether the subset has size at most k and spans all the elements. So the problem is in NP. To prove that it is $NP - complete$, we reduce Set Cover problem to this.

Given a collection of sets $S_1, S_2, \ldots, S_n \subset U$ such that $\cup_i S_i = U$ and some positive integer k, the Set Cover problem asks if there are at most k sets in the collection whose union is U. To reduce it to the spanoid rank problem, we can create a spanoid over $[n]$ elements where the inference rules are given by $A \models i$ iff $\cup_{j \in A} S_j \supset S_i$. The rank of this spanoid is at most k iff there exists k sets in the collection which cover all of U. ◀

3 Upper bounds on the rank of q-LCSs

In this section we prove the upper bounds on the rank of q-LCSs stated in Theorems 9 and 10. The proofs will rely on the set representation described in Section 2.2 and on random restriction and contraction arguments given below.

3.1 Graph theoretic lemmas

In this subsection, we will prove a key technical lemma about a random graph process that will be useful for proving upper bounds on the rank of q-LCSs. We denote by $\mathcal{D}(n)$ the set of simple directed graphs on n vertices. We always assume w.l.o.g that the set of vertices are the integers between 1 and n.

▶ **Definition 23** ((α, β)-spread distribution)**.** Let μ be a distribution on $\mathcal{D}(n)$. We say that μ is (α, β)-*spread* if the following conditions are true for a graph G sampled from μ:

1. Each vertex $i \in [n]$ has an incoming edge with probability at least α i.e.

$$\forall i \ \mathbf{Pr}_{G \sim \mu} [\exists j : (j, i) \in E(G)] \geq \alpha.$$

2. For every $i, j \in [n]$, the probability that (j, i) is an edge is at most β/n i.e.

$$\forall i, j \ \mathbf{Pr}_{G \sim \mu} [(j, i) \in E(G)] \leq \frac{\beta}{n}.$$

For example, one can generate an $(k/n, 1)$-spread distribution μ on $\mathcal{D}(n)$ in the following way: Fix arbitrary sets $S_1, \ldots, S_n \subset [n]$ of size k each. To sample a graph G from μ, pick a uniformly random element $j \in [n]$ and let G be the directed graph containing the edges (j, i) for each i such that $j \in S_i$. This satisfies the definition since for any fixed $i \in [n]$, i has an incoming edge if $j \in S_i$ which happens with probability $|S_i|/n = k/n$. And for any fixed $i', j' \in [n]$, the probability that (j', i') is an edge is at most $1/n$ since this happens only when $j' = j$ and j is chosen uniformly at random from $[n]$. Note that the sampled edges overall are highly correlated (they all have j as an endpoint).
We will need following simple observation about (α, β)-spread distributions.

▶ **Lemma 24.** *Let μ be an (α, β)-spread distribution on $\mathcal{D}(n)$. For every vertex i and every subset $S \subset [n]$ of size at most $\frac{\alpha n}{2\beta}$,*

$$\mathbf{Pr}_{G\sim\mu}[\exists j \notin S : (j,i) \in E(G)] \geq \frac{\alpha}{2}.$$

Proof. This follows from union bound and properties of (α, β)-spread distributions.

$$\begin{aligned}
\alpha &\leq \mathbf{Pr}_{G\sim\mu}[\exists j : (j,i) \in E(G)] \\
&\leq \mathbf{Pr}[\exists j \in S : (j,i) \in E(G)] + \mathbf{Pr}[\exists j \notin S : (j,i) \in E(G)] \\
&\leq \sum_{j\in S} \mathbf{Pr}[(j,i) \in E(G)] + \mathbf{Pr}[\exists j \notin S : (j,i) \in E(G)] \\
&\leq \frac{\alpha n}{2\beta} \cdot \frac{\beta}{n} + \mathbf{Pr}[\exists j \notin S : (j,i) \in E(G)] \\
&= \frac{\alpha}{2} + \mathbf{Pr}[\exists j \notin S : (j,i) \in E(G)]
\end{aligned}$$

◀

Given a distribution μ on graphs we would like to study the random process in which we, at each iteration, sample from μ and 'add' the edges we got to the graph obtained so far. For two graphs G and H on the same set of vertices, we denote by $G \cup H$ their set theoretic union (as a union of edges).

▶ **Definition 25** (Graph process associated with μ)**.** Let μ be a distribution on $\mathcal{D}(n)$. We define a sequence of random variables G_t^μ, $t = 0, 1, 2, \ldots$ as follows. G_0^μ is the empty graph on $[n]$ vertices. At each step $t \geq 1$ we sample a graph G according to μ (independently from all previous samples) and set $G_t = G_{t-1} \cup G$.

For a graph $G \in \mathcal{D}(n)$ and a vertex $i \in [n]$ we denote by $\mathsf{Rea}(i)$ the set of vertices that are reachable from i (via walking on directed edges). By convention, a vertex is always reachable from itself. Similarly, for a set of vertices $S \subset [n]$ we denote by $\mathsf{Rea}(S) = \cup_{i\in S}\mathsf{Rea}(i)$ the set of vertices reachable from some vertex in S. We denote the set of strongly connected components of G by $\Gamma(G)$. We denote by $C(i) \in \Gamma(G)$ the strongly connected component of G containing i. We say that $C \in \Gamma(G)$ is a *source* if C has no incoming edges from any vertex not in C.

▶ **Lemma 26.** *Let μ be an (α, β)-spread distribution on $\mathcal{D}(n)$ and let G_t^μ be its associated graph process. Then, for all $t \geq 0$, there is positive probability that the graph $\Gamma(G_t^\mu)$ has at most*

$$n \cdot (1 - \alpha/4)^t + \frac{2\beta}{\alpha}$$

sources.

Proof. If $C \in \Gamma(G)$ is a source, we define the *weight* of C to be the number of vertices reachable from C (including vertices of C) that are not reachable from any other source of G. More formally, let

$$\mathsf{Rea}'(C) = \{j \in \mathsf{Rea}(C) \mid j \notin \mathsf{Rea}(C'), \text{for all sources } C' \in \Gamma(G), C' \neq C\}.$$

Then the weight of a source $C \in \Gamma(C)$ is denoted by $w(C) = |\mathsf{Rea}'(C)|$ (we do not define weight for components that are not sources). Let us call a source $C \in \Gamma(G_t)$ 'light' if its weight $w(C)$ is at most $k = \frac{\alpha n}{2\beta}$ and 'heavy' otherwise. By the definition of weight, there could be at most $n/k = 2\beta/\alpha$ heavy sources.

We will argue that, in each step, as we move from G_t to G_{t+1}, the number of light sources must decrease by a factor of $(1-\alpha/4)$ with positive probability. For that purpose, suppose there are m_t sources in G_t and among them m'_t are light. Fix some light source and pick a representative vertex i from it. Since i is contained in a light source, $|\mathsf{Rea}'(C(i))| \leq \frac{\alpha n}{2\beta}$. When going to $G_{t+1} = G_t \cup G$, i gets an incoming edge from outside the set $\mathsf{Rea}'(C(i))$ with probability at least $\alpha/2$ by Lemma 24. If this happens then in G_{t+1}, this source will either stop being a source or merge with another source.

Picking a representative for each light source in G_t, we see that the expected number of representatives i which get a new incoming edge from outside $\mathsf{Rea}'(C(i))$ is at least $(\alpha/2)m'_t$. Hence, this quantity is obtained with positive probability. Now, if at least $(\alpha/2)m'_t$ light sources 'merge' with another source or stop being a source in G_{t+1} then the total number of light sources must decrease by at least $(\alpha/4)m'_t$ (the worst case being that $(\alpha/4)m'_t$ disjoint pairs of light sources merge with each other). Hence, with positive probability we get that $m'_{t+1} \leq m'_t \cdot (1-\alpha/4)$. Therefore, since the samples in each step t are independent, there is also a positive probability that $m'_t \leq n \cdot (1-\alpha/4)^t$ and $m_t \leq m'_t + 2\beta/\alpha$. This completes the proof. ◀

3.2 Proof of upper bound from Theorem 9

▶ **Theorem 27** (Rank of 2-LCSs). *Let $\mathcal{S}$ be a 2-LCS on $[n]$ with error-tolerance δ. Then $\mathsf{rank}(\mathcal{S}) \leq O(\frac{1}{\delta}\log_2 n)$.*

Proof. We will work with the (equivalent) set formulation: let $\mathcal{F} = \{S_1, \ldots, S_n\}$ be a set system representing the spanoid $\mathcal{S}$ as in Lemma 21.

We start by defining an (α, β)-spread distribution μ on $\mathcal{D}(n)$ as follows: To sample a graph G from μ we first pick $\ell \in [n]$ uniformly at random. Then we add a directed edge from j to i for every i, j such that $\{j, \ell\} \in M_i$. In this case we have $S_j \cap S_\ell \subseteq S_i$ and so, after restricting to S_ℓ we have $S_j \cap S_\ell \subseteq S_i \cap S_\ell$.

▶ **Claim 28.** *μ is a $(2\delta, 1)$-spread distribution.*

Proof. For any fixed $i \in [n]$, i will get an incoming edge if ℓ, which is randomly chosen from $[n]$, belongs to M_i. Since M_i has at least δn edges, this will happen with probability at least 2δ. Now fix any $i, j \in [n]$, (j, i) will be an edge iff ℓ is equal to the the vertex that matches j in the matching M_i, this happens with probability at most $1/n$. If j is not matched in M_i, the probability is zero. ◀

Consider the graph process G_t^μ and let $S_{\ell_1}, \ldots, S_{\ell_t}$ be the sets chosen in the t iterations of sampling from μ. If $i \in \mathsf{Rea}(j)$ in the graph G_t^μ, this means that, after restricting to the intersection $S = S_{\ell_1} \cap \ldots \cap S_{\ell_t}$, the set S_j is contained in S_i (i.e., $S_j \cap S \subseteq S_i \cap S$). By Lemma 26, after $t = O(\frac{1}{\delta}\log_2 n)$ steps, the graph process G_t^μ will contain $r = O(1/\delta)$ sources. Pick a representative $S_{a_1}, \ldots, S_{a_r}$ from each of these sources. Then, the intersection of the $t + r = O(\frac{1}{\delta}\log_2 n)$ sets $S_{\ell_1}, \ldots, S_{\ell_t}$ and $S_{a_1}, \ldots, S_{a_r}$ is contained in all n sets $S_1, \ldots, S_n$. That is because, when restricted to the intersection of $S_{\ell_1}, \ldots, S_{\ell_t}$, each set S_i contains one of the sets $S_{a_j}, j \in [r]$. ◀

3.3 Proof of upper bound from Theorem 10

▶ **Theorem 29** (Rank of q-query LCSs). *Let $\mathcal{S}$ be a q-LCS with error-tolerance δ and $q \geq 3$. Then*

$$\mathsf{rank}(\mathcal{S}) \leq O\left(\delta^{-\frac{1}{q-1}} \cdot n^{\frac{q-2}{q-1}} \log_2 n\right).$$

Proof. Like the 2-query case, we work with the set representation $\mathcal{F} = \{S_1, \ldots, S_n\}$ of $\mathcal{S}$ as in Lemma 21. We follow the same strategy as in the proof of the 2-query case. The difference is that, in this case, we will need to pick many sets to restrict to in each step instead of just one. The first observation is that, if $\{j_1, \ldots, j_q\} \in M_i$ then, restricted to the intersection $S = S_{j_1} \cap \ldots \cap S_{j_{q-1}}$ we have $S_{j_q} \subset S_i$. The second observation is that, if we choose a subset $J \subset [n]$ of size roughly $n^{\frac{q-2}{q-1}}$ then, in expectation, J will contain $q-1$ elements in one of the q-subsets of M_i for a constant fraction of the i's. Repeating this a logarithmic number of times and using Lemma 26, as in the proof of Theorem 27 will then complete the proof.

We start by defining an (α, β)-spread distribution μ on $\mathcal{D}(n)$. To sample a graph G from μ we first pick a random set $J \subset [n]$ such that each $j \in [n]$ is chosen to be in J independently with probability $(\delta n)^{-1/(q-1)}$. By Markov's inequality we have that

$$\mathbf{Pr}\left[|J| \geq 4 \cdot \delta^{-\frac{1}{q-1}} n^{\frac{q-2}{q-1}}\right] \leq 1/4. \tag{2}$$

For each $i \in [n]$ and each q-subset $T \in M_i$ we select $q-1$ elements of T arbitrarily and refer to them as the *distinguished* $(q-1)$*-subset of* T. We now argue that, for each $i \in [n]$, there is relatively high probability that J will contain the distinguished $(q-1)$-subset of at least one q-subset in M_i.

▶ **Claim 30.** *Let E_i denote the event that J contains the distinguished $(q-1)$-subset from at least one q-subset in M_i. Then, for each $i \in [n]$ we have that $\mathbf{Pr}[E_i] \geq 1/2$.*

Proof. J will contain the distinguished $q-1$ elements in a specific q-subset with probability $(\delta n)^{-1}$. Since the δn q-subsets in M_i are disjoint, the probability that J will not contain any of the distinguished $(q-1)$-subsets is at most $(1-(1/\delta n))^{\delta n} \leq 1/2$. ◀

We are now ready to define the edges in the graph G sampled by μ. First we check if $|J| \geq 4 \cdot \delta^{-\frac{1}{q-1}} n^{\frac{q-2}{q-1}}$. If this is the case then μ outputs the empty graph (by Eq.2 this happens with probability at most $1/4$). Otherwise for each $i \in [n]$ we check to see if J contains the distinguished $(q-1)$-subset from one of the q-subsets of M_i. If there is at least one such q-subset, we pick one of them uniformly at random. Suppose the q-subset we chose is $\{j_1, \ldots, j_q\}$ and that the distinguished elements are the first $q-1$. Then we add the directed edge $j_q \to i$ to the graph G. By the above discussion, we know that, restricted to the intersection of all sets indexed by J the set S_{j_q} is contained in S_i (hence the directed edge representing set inclusion).

▶ **Claim 31.** *μ is $(1/4, 1/\delta)$-spread.*

Proof. By Claim 30, and since the probability that $|J|$ is too large is at most $1/4$ we see that any fixed $i \in [n]$ will get an incoming edge in G with probability at least $\alpha = 1/4$. For any fixed $i, j \in [n]$, since the distribution of the special q-subset which is contributing an edge to i is uniform in M_i (conditioned on J containing a q-subset from M_i), we can conclude that $(j, i) \in E(G)$ with probability at most $1/(\delta n) = \beta/n$. This proves the claim. ◀

Now, applying Lemma 26, we get that, after $t = O(\log_2 n)$ steps, the graph process G_t^μ will contain at most $O(1/\delta)$ sources with positive probability. Let $J_1, \ldots, J_t$ be the sets chosen in the different steps of the process and, w.l.o.g, remove any of them that were too big (i.e., when the graph sampled by μ was empty). Hence, all of the sets satisfy $|J_i| \leq 4 \cdot \delta^{-\frac{1}{q-1}} n^{\frac{q-2}{q-1}}$. Now, let S be the intersection of all sets S_j such that j belongs to at least one of the sets J_i.

Then, restricted to S, each of the sets S_i contains one of the sources in the graph G_t^μ. Hence, if we add to our intersection a representative form each of the sources, we will get a set that is contained in all the sets S_j. The total number of sets we end up intersecting is bounded by

$$O(1/\delta) + \sum_{i=1}^{t} |J_i| = O\left(\delta^{-\frac{1}{q-1}} \cdot n^{\frac{q-2}{q-1}} \log_2 n\right).$$

This completes the proof of the theorem. ◀

4 Constructing q-LCSs with high rank

In this section we prove the lower bound part of Theorem 10 (the lower bound for the 2-query case follows from the Hadamard code construction). We will in fact generate this spanoid at random by picking, for each $i \in [n]$, a random q-matching M_i on $[n]$ and, for each q-subset $T \in M_i$ add the rule $T \models i$. The resulting spanoid will thus have, by design, the structure of a q-LCS. The reason why this spanoid should have high rank (with high probability) relies on the following observation. Suppose $A \subset [n]$ is a set that spans $[n]$. This means that there is a sequence of derivations $T_i \models i$ with each q-subset T_i in the matching M_i that eventually generates all of $[n]$. We can limit ourselves to the first $C \cdot |A|$ such derivations for some large C. These derivations generate a set A' of size $(C+1)|A|$ (including the original A and the $C|A|$ newly derived elements). Now, the set A' must contain all of the q-subsets T_i for $C|A|$ values of i. However, the union of randomly chosen $C|A|$ q-subsets will generally have size much larger than $(C+1)|A|$ (closer to $q \cdot C|A|$).

▶ **Theorem 32** (Existence of high rank q-LCSs). *For any integer $q \geq 3$ and all sufficiently large n the following holds. Consider the following distribution generating a spanoid $\mathcal{S}$ on base set $[n]$. For each $i \in [n]$ pick a q-matching M_i of size $\lfloor n/2q \rfloor$ uniformly at random and add the rule $T \models i$ for all $T \in M_i$. Then, with probability approaching one, $\mathsf{rank}(\mathcal{S})$ is larger than $r = cn^{\frac{q-1}{q-2}}/\log_2(n)$, where $0 < c < 1$ is an absolute constant.*

Proof. Let $m = r \cdot \log_2(n) = cn^{\frac{q-1}{q-2}}$. If the rank of $\mathcal{S}$ is at most r then there exists a set $A \subset [n]$ of size r that spans (using the rules obtained from the n random matchings $M_1, \ldots, M_n$) the entire base set $[n]$. We will upper bound the probability that such a set exists by bounding the smaller event given by the existence of a set of m rules that can be applied one after another starting with the original set A. That is, let $\mathcal{E}$ denote the event that there exists a set A of size r on which one can sequentially apply m rules of the form $T_{j_i} \models j_i$ with each T_{j_i} belonging to the matching M_{j_i} and for m different values $j_1, \ldots, j_m \in [n]$ arriving at the final set $\hat{A} = A \cup \{j_1, \ldots, j_m\}$. If A spans $[n]$ then clearly the event $\mathcal{E}$ must hold and so, it is enough to show that $\mathcal{E}$ happens with probability approaching zero.

We will present the event $\mathcal{E}$ as the union of (possibly overlapping) smaller events and then use the union bound, bounding the probability that each one occurs and multiplying by the number of bad events. Given a set $A \subset [n]$ of size r, a tuple of m indices $\hat{J} = \{j_1, j_2, \ldots, j_m\}$ and a family of q-subsets $\hat{T} = \{T_{j_1}, \ldots, T_{j_m}\}$ with $T_{j_i} \in M_{j_i}$ denote by $\mathcal{E}(A, \hat{J}, \hat{T})$ the event in which the set A spans the set $\hat{A} = A \cup \hat{J}$ using the rules $T_{j_i} \models j_i$ applied in order with i going from 1 to m. For every fixing of $A, \hat{J}, \hat{T}$ we can bound

$$\mathbf{Pr}[\mathcal{E}(A, \hat{J}, \hat{T})] \leq \prod_{i=1}^{m} \mathbf{Pr}[T_{j_i} \subset \hat{A}].$$

W.l.o.g suppose we sample the random matchings iteratively, picking a new q-subset at random among the available elements not covered by any previously chosen q-subsets in the current matching. Since the number of q-subsets in each matching is $\lfloor n/2q \rfloor$ we have, at each step, at least $n/2$ available elements to chose from and so

$$\mathbf{Pr}[T_{j_i} \subset \hat{A}] \leq \frac{\binom{m+r}{q}}{\binom{n/2}{q}} \leq \left(\frac{4m}{n}\right)^q.$$

Taking the product over all m q-subsets in $\hat{T}$ we get

$$\mathbf{Pr}[\mathcal{E}(A, \hat{J}, \hat{T})] \leq \left(\frac{4m}{n}\right)^{qm}.$$

To complete the proof we bound the number of tuples $(A, \hat{J}, \hat{T})$ as above by

$$\binom{n}{r} \cdot \binom{n}{m} \cdot \lfloor n/2q \rfloor^m \leq n^r \cdot (en/m)^m \cdot n^m \leq \left(\frac{6n^2}{m}\right)^m,$$

where the last inequality used the fact that $r/m \leq 1/\log_2(n)$. Putting these bounds together we get that

$$\mathbf{Pr}[\mathcal{E}] \leq \left(\frac{4m}{n}\right)^{qm} \left(\frac{6n^2}{m}\right)^m = \left(\frac{6 \cdot 4^q \cdot m^{q-1}}{n^{q-2}}\right)^m$$

which is exponentially decreasing in m for the given choice of $m = c \cdot n^{(q-2)/(q-1)}$ and for c a sufficiently small constant. ◀

One could ask for a more explicit construction of an LCS with rank equal to (or even close to) that stated above. We are not able to give such a construction but can relate this problem to a longstanding open problem in explicit construction of expander graphs. A bipartite (balanced) expander of degree q, is a bipartite graph with n left vertices L and n right vertices R such that the degree of each vertex is q and such that sets $A \subset L$ of size 'not too large' have many neighbors in R. More specifically, one typically asks that sets with $|A| \leq n/2$ have at least $(1+\epsilon)|A|$ right neighbors for some constant $\epsilon > 0$. It is quite easy to see that a random graph of this form will be a good expander with high probability and, by now, there are also many explicit constructions [17]. One can also consider *unbalanced* bipartite expanders in which $|L| \gg |R|$. Take, for example, the setting in which $|L| = n^2, |R| = n$ and when the degree of every vertex in L is some constant q. A simple probabilistic argument shows that sufficiently small sets in L, namely sets of size $|A| \leq n^{\alpha_q}$ with $\alpha_q < 1$ a constant depending on q and approaching 1 as q grows, have many neighbors in R (say, at least $2|A|$). However, no explicit constructions of such graphs are known (for any constant q and any $\alpha_q > 0$). The property we needed in our random construction of LCSs can be thought of as an 'easier' variant of the expander construction problem. Given q-matchings $M_1, \ldots, M_n$ each of size δn consider the bipartite graph with $L = [n] \times [\delta n]$ and $R = [n]$. We identify each vertex $(i, j) \in L$ with the j'th q-subset T_{ij} of M_i and connect it to the q neighbors in R given by that q-subset. For our proof to work we need the property that there is no small set containing many q-subsets from different matchings. This corresponds to asking for the above graph to be an expander for a restricted family of sets, namely to sets that have at most one vertex (i, j) for a given i (with each subgraph $(i, *)$ defining a matching).

5 Conclusion and open problems

Our work introduces the abstract notion of a spanoid in the hope that further study of its properties will lead to progress on LCCs and perhaps in other areas. We list below some concrete directions for future work.

1. We showed that there exist spanoids, called q-LCSs, which "look like" q-LCCs and whose rank matches the best known upper bounds. Can we bypass this 'barrier' by using additional properties of LCCs? We have at least two examples where this was possible. One is the result of [21] for LCCs over constant size alphabet and the other is the work in [11] for linear 3-LCCs over the real numbers. The bounds of [21] crucially depend on the alphabet having small size and the bounds in [11] exploit properties of real numbers.
2. Understanding the possible gap between functional rank and formal rank of a spanoid is a very interesting question. We proved that there can be a polynomial gap. The next challenge is to find a spanoid on n elements whose $\mathsf{f\text{-}rank}$ is $n^{o(1)}$ and rank is $n^{\Omega(1)}$. Naturally, q-LCSs for constant $q \geq 3$ are plausible candidates for this. If there are no such spanoids, then it would imply the existence of q-LCCs of length n and $n^{\Omega_q(1)}$ dimension![1]
3. Suppose we start with a functional representation with large alphabet, can we do alphabet reduction without losing too many codewords?
4. We have seen that one way to go past the rank barrier is to use $\mathsf{LP}^{entropy}$. Can we improve the existing upper bounds on the dimension of q-LCCs by upper bounding $\mathsf{LP}^{entropy}$ of q-LCSs? Can we use LP duality and construct good feasible solutions to the dual of $\mathsf{LP}^{entropy}$ to prove good upper bounds on $\mathsf{LP}^{entropy}$?
5. What are other connections of spanoids to existing theory of set systems, matroids, algebraic equations and other problems described in the introduction?

References

1 Michael Alekhnovich, Eli Ben-Sasson, Alexander A Razborov, and Avi Wigderson. Pseudorandom generators in propositional proof complexity. *SIAM Journal on Computing*, 34(1):67–88, 2004.

2 Michael Alekhnovich and Alexander A Razborov. Lower bounds for polynomial calculus: Non-binomial case. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 190–199. IEEE, 2001.

3 Boaz Barak, Zeev Dvir, Avi Wigderson, and Amir Yehudayoff. Fractional Sylvester-Gallai theorems. *Proceedings of the National Academy of Sciences*, 2012.

4 Boaz Barak, Zeev Dvir, Amir Yehudayoff, and Avi Wigderson. Rank bounds for design matrices with applications to combinatorial geometry and locally correctable codes. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 519–528. ACM, 2011.

5 Eli Ben-Sasson and Avi Wigderson. Short proofs are narrow—resolution made simple. *Journal of the ACM (JACM)*, 48(2):149–169, 2001.

6 Arnab Bhattacharyya, Sivakanth Gopi, and Avishay Tal. Lower Bounds for 2-Query LCCs over Large Alphabet. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, pages 30:1–30:20, 2017. doi:10.4230/LIPIcs.APPROX-RANDOM.2017.30.

7 Béla Bollobás. Weakly k-saturated graphs. Beiträge zur Graphentheorie (Kolloquium, Manebach, 1967), Teubner, Leipzig, pages 25–31, 1968.

[1] Possibly over a large alphabet.

8 Henning Bruhn and Oliver Schaudt. The journey of the union-closed sets conjecture. *Graphs and Combinatorics*, 31(6):2043–2074, 2015.

9 Irit Dinur and Tali Kaufman. Dense locally testable codes cannot have constant rate and distance. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 507–518. Springer, 2011.

10 Zeev Dvir. Incidence Theorems and Their Applications. *Foundations and Trends® in Theoretical Computer Science*, 6(4):257–393, 2012. doi:10.1561/0400000056.

11 Zeev Dvir, Shubhangi Saraf, and Avi Wigderson. Breaking the quadratic barrier for 3-LCC's over the reals. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 784–793. ACM, 2014.

12 Klim Efremenko. 3-query locally decodable codes of subexponential length. In *STOC*, pages 39–44, 2009. doi:10.1145/1536414.1536422.

13 Paul Erdös, Peter Frankl, and Zoltán Füredi. Families of finite sets in which no set is covered by the union ofr others. *Israel Journal of Mathematics*, 51(1):79–89, 1985.

14 Jacob Fox, Choongbum Lee, and Benny Sudakov. Maximum union-free subfamilies. *Israel Journal of Mathematics*, 191(2):959–971, 2012.

15 Zoltán Füredi. Onr-Cover-free Families. *J. Comb. Theory Ser. A*, 73(1):172–173, January 1996. doi:10.1006/jcta.1996.0012.

16 Lianna Hambardzumyan, Hamed Hatami, and Yingjie Qian. Polynomial method and graph bootstrap percolation. *arXiv preprint*, 2017. arXiv:1708.04640.

17 Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

18 Eran Iceland and Alex Samorodnitsky. On coset leader graphs of structured linear codes. *arXiv preprint*, 2018. arXiv:1802.01184.

19 Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proceedings of the 32nd annual ACM symposium on Theory of computing (STOC 2000)*, pages 80–86. ACM Press, 2000.

20 Leroy Milton Kelly. A resolution of the Sylvester-Gallai problem of J.-P. Serre. *Discrete & Computational Geometry*, 1(2):101–104, 1986.

21 Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *J. of Computer and System Sciences*, 69:395–420, 2004. Preliminary version appeared in STOC'03.

22 Emanuel Knill. Graph generated union-closed families of sets. *arXiv preprint*, 1994. arXiv:math/9409215.

23 Amir Shpilka. Sylvester-Gallai type theorems for quadratic polynomials. *Private communication*, 2018.

24 Piotr Wójcik. Union-closed families of sets. *Discrete Mathematics*, 199(1-3):173–182, 1999.

25 David Woodruff. New Lower Bounds for General Locally Decodable Codes. *Electronic Colloquium on Computational Complexity (ECCC)*, 14(006), 2007.

26 Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. *Journal of the ACM (JACM)*, 55(1):1, 2008.

27 Sergey Yekhanin. Locally decodable codes. *Foundations and Trends® in Theoretical Computer Science*, 6(3):139–255, 2012.