



INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus

Mathématique

Matthew Kwan, Ashwin Sah and Mehtaab Sawhney

Enumerating Matroids and Linear Spaces

Volume 361 (2023), p. 565-575

Published online: 1 February 2023

<https://doi.org/10.5802/crmath.423>



This article is licensed under the
CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.
<http://creativecommons.org/licenses/by/4.0/>



Les Comptes Rendus. Mathématique sont membres du
Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 1778-3569



Combinatorics / *Combinatoire*

Enumerating Matroids and Linear Spaces

Matthew Kwan^a, Ashwin Sah^b and Mehtaab Sawhney^{*, b}

^a Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria

^b Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

E-mails: matthew.kwan@ist.ac.at (M. Kwan), asah@mit.edu (A. Sah), msawhney@mit.edu (M. Sawhney)

Abstract. We show that the number of linear spaces on a set of n points and the number of rank-3 matroids on a ground set of size n are both of the form $(cn + o(n))n^{2/6}$, where $c = e^{\sqrt{3}/2-3}(1 + \sqrt{3})/2$. This is the final piece of the puzzle for enumerating fixed-rank matroids at this level of accuracy: there are exact formulas for enumeration of rank-1 and rank-2 matroids, and it was recently proved by van der Hofstad, Pendavingh, and van der Pol that for constant $r \geq 4$ there are $(e^{1-r}n + o(n))n^{r-1/r!}$ rank- r matroids on a ground set of size n .

Funding. Sah and Sawhney were supported by NSF Graduate Research Fellowship Program DGE-1745302. Sah was supported by the PD Soros Fellowship.

Manuscript received 14 December 2021, accepted 12 September 2022.

1. Introduction

Matroids (also sometimes known as *combinatorial geometries*) are fundamental objects that abstract the combinatorial properties of linear independence in vector spaces. Specifically, a matroid consists of a ground set E and a collection \mathcal{I} of subsets of E called independent sets¹. The defining properties of a matroid are that:

- the empty set is independent (that is, $\emptyset \in \mathcal{I}$);
- subsets of independent sets are independent (if $A' \subseteq A \subseteq E$ and $A \in \mathcal{I}$, then $A' \in \mathcal{I}$);
- if A and B are independent sets, and $|A| > |B|$, then an independent set can be constructed by adding an element of $A \setminus B$ to B (there is $a \in A \setminus B$ such that $B \cup \{a\} \in \mathcal{I}$).

Observe that any finite set of elements in a vector space (over any field) naturally gives rise to a matroid, though most matroids do not arise this way. The *rank* of a matroid is the maximum size of an independent set.

* Corresponding author.

¹Instead of defining a matroid by its collection of independent sets, some authors prefer to define a matroid by some other (equivalent) data, such as its collection of *flats*, its collection of *hyperplanes*, or its *rank function*. See for example [14, 20] for a more thorough introduction to matroids and their various definitions.

Enumeration of matroids is a classical topic, though the state of our knowledge is rather incomplete. Some early upper and lower bounds on the total number of matroids on a ground set of size n were obtained in the 1970s by Piff and Welsh [17], Piff [16] and Knuth [9], and these bounds were improved only recently by Bansal, Pendavingh, and van der Pol [2]. It is also of interest to enumerate matroids of fixed rank: let $m(n, r)$ be the number of rank- r matroids on a ground set of size n . It is trivial to see that $m(n, 1) = 2^n - 1$, and it is also possible to prove the exact identity $m(n, 2) = b(n + 1) - 2^n$, where $b(m)$ is the m th Bell number (which counts the number of partitions of an m -element set). This identity seems to have been first observed by Acketa [1].

For $r \geq 3$, an exact expression for $m(n, r)$ in terms of well-known functions does not seem to be possible², but after some exciting recent developments, rather precise asymptotic expressions have become available. First, Pendavingh and van der Pol [15] observed that (for constant $r \geq 1$) the lower bound $m(n, r) \geq (e^{1-r}n + o(n))^{n^{r-1}/r!}$ follows from Keevash's breakthrough work [7, 8] on existence and enumeration of *combinatorial designs*. They also proved an upper bound of the form $m(n, r) \leq (en + o(n))^{n^{r-1}/r!}$. Even more recently, van der Hofstad, Pendavingh and van der Pol [5] closed the gap for all $r \neq 3$, proving that $m(n, r) = (e^{1-r}n + o(n))^{n^{r-1}/r!}$ for constant $r \geq 4$. In the remaining case $r = 3$ they were able to prove $m(n, 3) \leq (ne^{1+\beta} + o(n))^{n^2/6} \approx (1.4n)^{n^2/6}$, where $-0.67 < \beta < -0.65$ is the solution to a certain variational problem. In this paper, we close the gap completely in this case $r = 3$.

Theorem 1.

$$m(n, 3) = \left(\frac{1 + \sqrt{3}}{2} e^{\sqrt{3}/2 - 3} n + o(n) \right)^{n^2/6} \approx (0.16169n)^{n^2/6}.$$

In particular, Theorem 1 shows that the lower bound $m(n, 3) \geq (e^{-2}n + o(n))^{n^2/6}$ obtainable from Keevash's results is far from sharp. This confirms a conjecture in [5] (and disproves the earlier [18, Conjecture 8.2.9]).

In fact, Theorem 1 is really a corollary of the following theorem, estimating the number of *linear spaces* on a set of n points. In incidence geometry, a linear space on a point set P is a collection of subsets of at least two points of P (called *lines*) such that each pair of points lies in a unique line (see for example [3, 19] for more on linear spaces). For reasons that will become clear in a moment, we denote the number of linear spaces on a set of n points by $p(n, 3)$.

Theorem 2.

$$p(n, 3) = \left(\frac{1 + \sqrt{3}}{2} e^{\sqrt{3}/2 - 3} n + o(n) \right)^{n^2/6} \approx (0.16169n)^{n^2/6}.$$

We remark that one may also be interested in linear spaces in which no line has exactly 2 points (these are called *proper* linear spaces). It should be possible to adapt our proof to show that the expression in Theorem 2 is also a valid estimate for the number of proper linear spaces on a set of n points (though this would require some rather deep machinery due to Keevash [8] and McKay and Wormald [12]). See Remark 10 for discussion.

To explain the connection between Theorems 2 and 1 we need to make a few more definitions. A d -*partition* (or *generalised partition of type d*) of a ground set E is a collection of subsets of E (called *parts*) each having size at least d , such that every subset of d elements of E is contained in exactly one of the parts. So, a 1-partition is an ordinary partition, and a 2-partition is a linear space. For any $r \geq 2$, there is a correspondence between the set of $(r - 1)$ -partitions of E and the set of so-called *paving matroids* of rank r on the ground set E . Namely, a paving matroid of rank

²Though, a lot of computational work has been done for small n, r , and there are many conjectures about the relations between the different $m(n, r)$; see for example [4] and the index for matroids on the On-Line Encyclopedia of Integer Sequences [13].

r is a matroid for which its set of *hyperplanes* (maximal subsets with rank $r - 1$) form an $(r - 1)$ -partition of its ground set. See for example [14, Section 3] for more details.

For $r \geq 2$ let $p(n, r)$ be the number of paving matroids of rank r , or equivalently the number of $(r - 1)$ -partitions, on a ground set of size n . Given the above correspondence, we trivially have $p(n, r) \leq m(n, r)$, and it was proved by Pendavingh and van der Pol [15, Theorem 3] that $p(n, r) \leq m(n, r) \leq p(n, r)^{1+O(1/n)}$ for constant r . So Theorem 1 is a direct consequence of Theorem 2, and for the rest of the paper we will abandon the language of matroids and focus on Theorem 2.

In fact, we find it convenient to use the language of graph theory: note that a linear space on a set of n points is precisely equivalent to a *clique-decomposition* of the complete graph K_n (meaning, a decomposition of the edges of K_n into nonempty cliques of arbitrary sizes).

1.1. Discussion of proof techniques

If one is interested in counting the number of decompositions of K_n into cliques which each have a *fixed* number of vertices k , this is a problem about enumerating *combinatorial designs*. Specifically, such a decomposition corresponds exactly to a design called a $(2, k, n)$ -Steiner system. Such Steiner systems can be enumerated using powerful tools due to Keevash [7, 8]: in particular, if n satisfies certain necessary divisibility conditions, the number of such Steiner systems can be written as $(c_k n + o(n))^{\alpha_k n^2}$, where $c_k^{k-2} = e^{1-\binom{k}{2}}/(k-2)!$ and $\alpha_k = (k-2)/(k(k-1))$. Note that α_k is maximised for two different k : namely, when $k = 3$ and when $k = 4$. This suggests that decompositions containing mostly 3-cliques and 4-cliques comprise the bulk of the clique-decompositions in $p(n, 3)$.

The above observation motivates our proof strategy, and we believe it also explains why counting $(r - 1)$ -partitions and rank- r matroids is most difficult when $r = 3$ (if $r \geq 4$, then one can do a similar calculation for *hypergraph* clique-decompositions and see that there is a single maximising value of k).

For the lower bound in Theorem 2 (namely, that there are *at least* about $(0.16n)^{n^2/6}$ clique-decompositions), we proceed in a very similar fashion as in [7]: we consider a random process that builds a clique-decomposition by iteratively removing random 3-cliques and 4-cliques from K_n (with a particular carefully chosen ratio between the two), until a very small number of edges remain (these edges are then treated as 2-cliques in our decomposition). We then study the number of possible outcomes of this process³. The details of the lower bound appear in Section 3.

The upper bound is more interesting. In [7], Keevash is able to upper-bound the number of Steiner systems by adapting an approach of Linial and Luria [11], using the so-called *entropy method*. Roughly speaking, the idea is as follows. To prove an upper bound on the number of k -clique decompositions of K_n , it suffices to prove an upper bound on the entropy of a uniformly random k -clique decomposition P . In order to specify an outcome of P , it suffices to specify, for each edge $e \in K_n$, the clique C_e containing e . Therefore, one can upper-bound the entropy of P by considering an ordering $e_1, \dots, e_{\binom{n}{2}}$ of the edges of K_n , and upper-bounding the conditional entropy of each C_{e_i} , given the previous cliques $C_{e_1}, \dots, C_{e_{i-1}}$. If $e_1, \dots, e_{\binom{n}{2}}$ is a *random* ordering, then it is possible to upper-bound these conditional entropies by studying the expected number of possible choices for C_{e_i} given $C_{e_1}, \dots, C_{e_{i-1}}$. This is possible due to a certain symmetry of k -clique decompositions: namely, Keevash makes crucial use of the fact that in *any* k -clique decomposition, for any edge e and almost all k -cliques $C \subseteq K_n$ containing e there are exactly

³It would be possible to consider a random process that, at each step, randomly decides whether to remove a 3-clique or 4-clique, with an appropriate probability. This would be in close correspondence with the upper bound approach described below. However, it is more convenient for us to reuse existing analysis of clique removal processes, and consider the concatenation of a 4-clique removal process and a 3-clique removal process.

$\binom{k}{2} - 1)^2$ edges $e' \notin C$ such that $C_{e'}$ and C share an edge (meaning that after $C_{e'}$ is revealed, C can be ruled out as a possible outcome of $C_{e'}$).

For decompositions of K_n into cliques of mixed sizes, an analogous symmetry property does not hold, and the number of edges e' whose clique $C_{e'}$ intersects a particular clique C depends on the structure of our clique-partition. So, we cannot prove the upper bound in Theorem 2 by a straightforward generalisation of Keevash's proof. Instead, we exploit a different symmetry property of clique-decompositions, generalising an observation in [10], as follows. Suppose P is a clique-partition into cliques of bounded size (say, each of the cliques in P has at most 100 vertices). Then, if we take the union of a *random* subset of the cliques in P , where each clique is included independently with probability $p \in (0, 1)$, we are very likely to arrive at a *quasirandom* graph with density about p (i.e., a graph whose "local statistics" resemble a random subgraph of K_n obtained by including each edge with probability p independently). Sweeping some details under the rug, this means that we can give an upper bound on the number of ways to choose a clique-decomposition with a prescribed number of cliques of each size (the precise statement is in Lemma 4), by counting in a clique-by-clique manner, where at each step the number of choices for a k -clique is roughly the expected number of k -cliques in a random graph of the appropriate density⁴. We remark that our approach seems to be more flexible than the entropy method, for problems of this type: it is possible to recover all of Keevash's upper bounds in this way (though with slightly weaker quantitative aspects). Also, in our view, our clique-by-clique approach is more naturally in correspondence with the clique-by-clique processes used to prove lower bounds in this area.

Finally, having an upper bound for the number of clique partitions with a prescribed number s_k of k -cliques for each $k \leq 100$ (and no cliques with more than 100 vertices), it remains to show that the contribution from cliques with more than 100 vertices is negligible, and to optimise our formula over choices of s_1, \dots, s_{100} . For the former, we use a very crude encoding argument (Lemma 3). The latter is a simple calculus exercise (essentially, we use the method of Lagrange multipliers; see Lemma 7). In agreement with the heuristic mentioned earlier, we find that our formula is maximised when only s_3, s_4 are non-negligible.

1.2. Notation

We use standard asymptotic notation throughout, as follows. For functions $f = f(n)$ and $g = g(n)$, we write $f = O(g)$ to mean that there is a constant C such that $|f| \leq C|g|$, $f = \Omega(g)$ to mean that there is a constant $c > 0$ such that $f(n) \geq c|g(n)|$ for sufficiently large n , and $f = o(g)$ to mean that $f/g \rightarrow 0$ as $n \rightarrow \infty$. Also, following [8], the notation $f = 1 \pm \varepsilon$ means $1 - \varepsilon \leq f \leq 1 + \varepsilon$.

We write $N_G(v)$ to denote the neighbourhood of a vertex v in a graph G (i.e., the set of vertices adjacent to v). For a real number x , the floor and ceiling functions are denoted $\lfloor x \rfloor = \max\{i \in \mathbb{Z} : i \leq x\}$ and $\lceil x \rceil = \min\{i \in \mathbb{Z} : i \geq x\}$. We will however mostly omit floor and ceiling symbols and assume large numbers are integers, wherever divisibility considerations are not important. Finally, all logarithms in this paper are in base e .

Acknowledgements

We thank Michael Simkin for helpful comments on the manuscript. We thank Zach Hunter for several corrections.

⁴In our actual proof, due to the fact that we are considering decompositions into cliques of mixed sizes, it is more convenient to consider small "chunks" of cliques with a representative number of cliques of each size, and estimate the number of choices for each chunk, rather than estimating the number of choices for each clique individually.

2. The upper bound

2.1. Removing the contribution from large parts

We first reduce to the case where all cliques have bounded size. Related ideas appeared in [5].

Lemma 3. *Fix $L \geq 11$ and n sufficiently large as a function of L . Let $\Gamma_{s_2, \dots, s_L; E}$ denote the set of clique-decompositions of K_n , for which there are E edges covered by cliques with more than L vertices, and there are s_k cliques with k vertices for each $2 \leq k \leq L$. Then*

$$|\Gamma_{s_2, \dots, s_L; E}| \leq n^{|E|/5} |\Gamma_{s_2+E, s_3, \dots, s_L; 0}|.$$

Proof. Fix a clique decomposition $P \in \Gamma_{s_2, \dots, s_L; E}$. Let P_1 be the “truncated” clique decomposition obtained from P by first removing each clique with more than L vertices, and then adding two-vertex cliques (i.e., single edges) for each of the edges which are no longer covered by a clique. Then P is uniquely determined by the pair (P_1, P_2) , where P_2 contains all the cliques in P with more than L vertices.

There are at most $2^{|E|-1}$ ways to choose a sequence s_{L+1}, \dots, s_n such that $\sum_{t=L+1}^n \binom{t}{2} s_t = |E|$. Indeed, such a sequence can be interpreted as an integer partition⁵ of $|E|$ (where we are only allowed to use parts which have size of the form $\binom{t}{2}$ for $t > L$). For each such s_{L+1}, \dots, s_n , the number of possibilities for P_2 which contain exactly s_t cliques of each size $t > L$ is at most $\prod_{t=L+1}^n (n^t)^{s_t} = n^\Sigma$, where

$$\Sigma = \sum_{t=L+1}^n t s_t \leq \frac{2}{L} \sum_{t=L+1}^n \binom{t}{2} s_t = \frac{2|E|}{L}.$$

Then, we observe that $2^{|E|-1} n^\Sigma \leq n^{|E|/5}$ for $L \geq 11$ and n large enough. □

2.2. Counting decompositions into prescribed numbers of bounded-size cliques

We now estimate the number of clique-decompositions with prescribed numbers of cliques of each (bounded) size. We will later optimise over choices of these prescribed numbers.

Lemma 4. *Fix a constant $L \in \mathbb{N}$ and integers s_2, \dots, s_L such that $\binom{2}{2} s_2 + \dots + \binom{L}{2} s_L = \binom{n}{2}$. Let Γ_{s_2, \dots, s_L} be the set of all clique-decompositions of K_n whose number of t -cliques is s_t for each t . Then*

$$|\Gamma_{s_2, \dots, s_L}| \leq \exp \left(\sum_{k=2}^L s_k \left(\log \binom{n}{k} - \log s_k + 1 \right) - \binom{n}{2} \pm n^{2-\Omega(1)} \right).$$

We define an *ordered* clique-decomposition of K_n to be an ordered list of cliques whose edge-disjoint union is equal to K_n . Let Ξ_{s_2, \dots, s_L} be the set of all orderings of clique-decompositions in Γ_{s_2, \dots, s_L} . First, we need the following modification of [10, Lemma 2.6], showing that for initial segments of a random ordered clique-decomposition, the graph of uncovered edges is “typical”/“quasirandom”.

Lemma 5. *Fix a constant $L \in \mathbb{N}$ and any integers s_2, \dots, s_L such that $\binom{2}{2} s_2 + \dots + \binom{L}{2} s_L = \binom{n}{2}$. Consider a uniformly random ordered clique-decomposition (of K_n) from Ξ_{s_2, \dots, s_L} (which has $N := s_2 + \dots + s_L$ cliques). Let G_m be the random subgraph of K_n consisting of the edges not appearing in the first m cliques of our random ordered clique-decomposition. Then with probability $1 - o_{n \rightarrow \infty}(1)$, for all $0 \leq m \leq N$ and all sets of vertices A with $|A| \leq L$, we have*

$$\left| \left| \bigcap_{w \in A} N_{G_m}(w) \right| - (1 - m/N)^{|A|} n \right| \leq n^{1/2} \log n.$$

⁵The number of partitions of an integer N is at most its number of compositions, which is 2^{N-1} .

Proof. Fix a particular choice of m and A ; we will take a union bound over all such choices. It suffices to consider a uniformly random ordering of a *fixed* clique-decomposition $P \in \Gamma_{s_2, \dots, s_L}$ (i.e., we prove the desired statement conditioned on any outcome of the unordered set of cliques in our random ordered clique-decomposition).

The first m cliques in our random ordering comprise a uniformly random subset $R \subseteq P$ of m cliques in P . Consider the closely related “binomial” random subset $R' \subseteq P$, where each clique is included with probability $1 - m/N$ independently; let G'_m contain the edges of K_n which do not appear in any of the cliques in R' .

Since the cliques in P are edge-disjoint, note that there are at most $\binom{|A|}{2} = O(1)$ cliques in P that include more than one vertex in A . Let U be the set of vertices in these atypical cliques. Now, for each $v \notin A$ and $w \in A$ there is exactly one clique e_v^w in P containing v and w , whose presence in R' would prevent v from appearing in $\bigcap_{v \in A} N_{G'_m}(v)$. For each fixed $v \notin U$ the hyperedges e_v^w , for $w \in A$, are distinct, so

$$\Pr\left(v \in \bigcap_{w \in A} N_{G'_m}(w)\right) = (1 - m/N)^{|A|}.$$

Let $Q = |\bigcap_{w \in A} N_{G'_m}(w)|$; it follows that $\mathbb{E}Q = (1 - m/N)^{|A|}n + O(1)$.

Now let I be the set of cliques of P which contain a vertex of A . We have $|I| = O(n)$ since the cliques in P are edge-disjoint. Note that Q is entirely determined by $I \cap R'$, and adding or removing any clique from R' affects Q by at most $L - 1 = O(1)$. So by the Azuma–Hoeffding inequality, for $t \geq 0$ we have

$$\Pr(|Q - \mathbb{E}Q| \geq t) \leq 2 \exp(-\Omega(t^2/n)).$$

It follows that with probability at least $1 - n^{-10L}$ we have $|Q - (1 - m/N)^{|A|}n| \leq \sqrt{n} \log n$. Recall that we have been considering the “binomial” random subset R' ; we can transfer this result to the “uniform” random subset R using a standard inequality (for example, the so-called Pittel inequality; see [6, p. 17]). Then, we take a union bound over choices of m and A . □

We also need the fact that the cliques of different sizes are “well-distributed” in a random ordered clique-decomposition.

Lemma 6. Fix a constant $L \in \mathbb{N}$ and integers s_2, \dots, s_L such that $\binom{s_2}{2} + \dots + \binom{s_L}{2} = \binom{n}{2}$. Consider a uniformly random ordered clique-decomposition (of K_n) from Ξ_{s_2, \dots, s_L} (which has $N := s_2 + \dots + s_L$ cliques). Then with probability $1 - o_{n \rightarrow \infty}(1)$, for any $0 \leq m < m' \leq N$ and any $0 \leq k \leq L$, if we consider all the cliques ranging from the $(m + 1)^{\text{th}}$ to the m'^{th} in our random ordered clique-decomposition, the number of such cliques that have exactly k vertices differs from $s_k(m' - m)/N$ by at most $n \log n$.

Proof. As in the proof of Lemma 5, it suffices to consider a uniformly random reordering of a *fixed* clique-decomposition $P \in \Xi_{s_2, \dots, s_L}$. The desired result then follows from a concentration inequality for the hypergeometric distribution (see for example [6, (2.5) and Theorem 2.10]) and the union bound. □

Now we are ready to prove Lemma 4.

Proof of Lemma 4. Let $N = s_2 + \dots + s_L$, and let c be a very small constant ($c = 1/(10L^2)$ will do). We will count *ordered* clique decompositions in Ξ_{s_2, \dots, s_L} , and then at the end of the proof we will divide by $N!$.

Partition the interval $\{1, \dots, N\}$ into sub-intervals I_1, \dots, I_{n^c} by taking

$$I_i = \{1, \dots, N\} \cap ((i - 1)Nn^{-c}, iNn^{-c}].$$

Let $m_i = \min I_i = \lfloor (i - 1)Nn^{-c} + 1 \rfloor$ be the first index in each I_i . Say that an ordered clique decomposition $P \in \Xi_{s_2, \dots, s_L}$ is *ordinary* if for each $1 \leq i \leq n^c$, the following hold.

- (1) The graph $G^{(i)} := G_{m_i-1}$ consisting of those edges not covered by the first $m_i - 1$ cliques of P satisfies the conclusion of Lemma 5.
- (2) For each $1 \leq i \leq n^c$ and $2 \leq k \leq L$, the number of cliques ranging from the m_i^{th} to the $(m_{i+1} - 1)^{\text{th}}$ which have exactly k vertices satisfies the conclusion of Lemma 6.

Almost all ordered clique decompositions in Ξ_{s_2, \dots, s_L} are ordinary by Lemmas 5 and 6, so it suffices to prove an upper bound on the number of ordinary decompositions.

For each $1 \leq i \leq n^c$, we consider separately the number of choices for the cliques indexed by indices in I_i , for an ordinary ordered clique-decomposition. Let $\gamma_i = 1 - (i - 1)n^{-c}$. Now, (1) implies that for all $k \leq L$, the number of k -cliques in $G^{(i)}$ is

$$\gamma_i^{\binom{k}{2}} n^k / k! + O\left(n^{k-1/2} \log n\right). \tag{1}$$

To see this, we count the number of ways to choose an ordered list of k vertices inducing a clique, in a vertex-by-vertex fashion, then divide by $k!$.

For any choice of $t_k = n^{-c} s_k + O(n \log n)$, we have

$$\binom{t_2 + \dots + t_L}{t_2, \dots, t_L} = e^{O(n(\log n)^2)} \left(\prod_{k=2}^L \left(\frac{s_k}{s_2 + \dots + s_k} \right)^{-\frac{s_k}{s_2 + \dots + s_k}} \right)^{n^{-c}(s_2 + \dots + s_k) + O(n \log n)}.$$

So, given (2), we can multiply these estimates to see that the number of ways to choose the cliques indexed by I_i is at most

$$e^{n^{2-c-\Omega(1)}} \prod_{k=2}^L \left(\frac{s_k}{N} \right)^{-s_k n^{-c}} \prod_{k=2}^L \left(\gamma_i^{\binom{k}{2}} \frac{n^k}{k!} \right)^{n^{-c} s_k}.$$

We next take the product of this expression over all $1 \leq i \leq n^c$, and divide by the number of orderings $N!$ of each clique-decomposition, to obtain the desired result. In particular one obtains

$$\begin{aligned} & e^{n^{2-\Omega(1)}} \frac{1}{N!} \prod_{i=1}^{n^c} \left(\prod_{k=2}^L \left(\frac{s_k}{N} \right)^{-s_k n^{-c}} \prod_{t=2}^L \left(\gamma_i^{\binom{k}{2}} \frac{n^k}{k!} \right)^{n^{-c} s_k} \right) \\ &= e^{n^{2-\Omega(1)}} \frac{1}{N!} \prod_{k=2}^L \left(\left(\frac{s_k}{N} \right)^{-s_k} \left(\frac{n^k}{k!} \right)^{s_k} \right) \cdot \prod_{i=1}^{n^c} \prod_{t=2}^L \gamma_i^{n^{-c} \binom{k}{2} s_k} \\ &= e^{n^{2-\Omega(1)} - \binom{n}{2}} \prod_{k=2}^L \left(\frac{s_k}{e} \right)^{-s_k} \left(\frac{n^k}{k!} \right)^{s_k}. \end{aligned}$$

We note that this involves an approximation by a Riemann integral:

$$\begin{aligned} \prod_{i=1}^{n^c} \prod_{k=2}^L \gamma_i^{n^{-c} s_k} &= \exp \left(n^{-c} \sum_{i=1}^{n^c} \log \gamma_i \sum_{k=2}^L \binom{k}{2} s_k \right) = \exp \left(\binom{n}{2} n^{-c} \sum_{i=1}^{n^c} \log (i n^{-c}) \right) \\ &= \exp \left(\binom{n}{2} \left(\int_0^1 \log x \, dx \pm O(n^{-c/2}) \right) \right) \\ &= \exp \left(-\binom{n}{2} + O(n^{2-c/2}) \right). \quad \square \end{aligned}$$

2.3. Optimising over prescribed clique numbers

Given Lemmas 3 and 4, the upper bound in Theorem 2 will be a simple consequence of the following lemma.

Lemma 7. Fix a constant $L \in \mathbb{N}$, let D be the set of $(s_2, \dots, s_L) \in \mathbb{R}^{L-1}$ such that $s_2, \dots, s_L \geq 0$ and $\binom{2}{2}s_2 + \dots + \binom{L}{2}s_L = \binom{n}{2}$, and consider the real-valued function $f : D \rightarrow \mathbb{R}$ defined by

$$f(s_2, \dots, s_L) = \frac{1}{5} s_2 \log n + \sum_{k=2}^L \left(s_k \log \binom{n}{k} - s_k \log s_k + s_k \right) - \binom{n}{2}.$$

Then, the maximum value of $f(s_2, \dots, s_L)$ is

$$\frac{n^2}{6} \left(\log n - 3 + \frac{\sqrt{3}}{2} + \log \frac{1 + \sqrt{3}}{2} \pm n^{-\Omega(1)} \right).$$

Proof. Since f is continuous on D and D is compact, our function f attains a maximum.

First, we claim that a maximum can only be attained when all s_k are strictly positive. Indeed, consider some $(s_1, \dots, s_L) \in D$ for which $s_k = 0$, in which case there must be some $s_j > 0$. We make a slight perturbation: increase s_k to $\binom{j}{2}\varepsilon$ and decrease s_j by $\binom{k}{2}\varepsilon$, for some very small $\varepsilon > 0$ (note that we are still in D), and consider the corresponding change to the value of f . Note that the terms containing s_j decrease by $O(\varepsilon)$ but the terms containing s_k increase by $\Omega(\varepsilon \log(1/\varepsilon))$. So, our perturbation has increased the value of f , which proves the claim.

Note that if we increase s_k by $\binom{j}{2}\varepsilon$ and decrease s_j by $-\binom{k}{2}\varepsilon$, for some very small $\varepsilon > 0$, then the value of f increases by

$$\binom{j}{2} \left(\log \binom{n}{k} - \log s_k + \tau(k) \right) \varepsilon - \binom{k}{2} \left(\log \binom{n}{j} - \log s_j + \tau(j) \right) \varepsilon + O(\varepsilon^2),$$

where we set $\tau(k) = (\log n)/5$ when $k = 2$, and $\tau(k) = 0$ when $k \neq 2$. (This essentially follows from taking a derivative). So, a maximum can only occur when each

$$\frac{\log \binom{n}{k} - \log s_k + \tau(k)}{\binom{k}{2}}$$

takes a common value λ . For this λ , we see that

$$s_k = \binom{n}{k} e^{-\binom{k}{2}\lambda + \tau(k)}. \tag{2}$$

Now, recalling the definition of D , we have

$$n^{1/5} \binom{n}{2} e^{-\lambda} + \sum_{k=3}^L \binom{k}{2} \binom{n}{k} e^{-\binom{k}{2}\lambda} = \binom{n}{2}. \tag{3}$$

There is a unique λ satisfying this equation, because the left-hand side of the equation is monotonically decreasing in λ . Now, if $\lambda = (\log n)/3 + \alpha$, for $|\alpha| \leq 1$, then we may compute

$$n^{1/5} \binom{n}{2} e^{-\lambda} + \sum_{k=3}^L \binom{k}{2} \binom{n}{k} e^{-\binom{k}{2}\lambda} = (e^{-3\alpha}/2 + e^{-6\alpha}/4) n^2 \pm n^{2-\Omega(1)}.$$

(The dominant terms of the expression on the left-hand side are the ones with $k \in \{3, 4\}$). Therefore, we can write the solution to Equation (3) as $\lambda = (\log n)/3 + \alpha$ for $|\alpha| \leq 1$ (since $\alpha = -1$ makes the left side of Equation (3) too large and $\alpha = 1$ makes it too small).

Recall that $\binom{n}{2} = (1/2)n^2 + O(n)$, so for the λ satisfying Equation (3) we have

$$e^{-3\alpha}/2 + e^{-6\alpha}/4 = 1/2 + n^{-\Omega(1)},$$

hence the quadratic formula yields

$$e^{-3\alpha} = \sqrt{3} - 1 + n^{-\Omega(1)}.$$

Thus

$$e^{-3\lambda} = (\sqrt{3} - 1) n^{-1} \pm n^{-1-\Omega(1)}.$$

Using Equation (2), we then compute that f is maximised when

$$s_3 = \frac{n^3}{6} e^{-3\lambda} \pm n^{2-\Omega(1)} = \frac{n^2(\sqrt{3}-1)}{6} \pm n^{2-\Omega(1)},$$

$$s_4 = \frac{n^4}{24} e^{-6\lambda} \pm n^{2-\Omega(1)} = \frac{n^2(2-\sqrt{3})}{12} \pm n^{2-\Omega(1)},$$

and $s_k = n^{2-\Omega(1)}$ for $k \notin \{3, 4\}$. The desired result follows after substituting into the formula for $f(s_2, \dots, s_L)$ and simplifying. \square

2.4. Deducing the upper bound

We now give the short deduction of the upper bound in Theorem 2 using Lemmas 3, 4 and 7.

Proof of the upper bound in Theorem 2. Let $L = 11$. The sets $\Gamma_{s_2, \dots, s_L, E}$ defined in Lemma 3 form a partition of the set of all clique-decompositions of K_n . There are at most $\binom{n}{2}^L = e^{n^{2-\Omega(1)}}$ choices of s_2, \dots, s_L, E , so it suffices to upper-bound the maximum possible value of $|\Gamma_{s_2, \dots, s_L, E}|$. By Lemma 3 it in fact suffices to upper-bound the maximum possible value of $n^{s_2/5} |\Gamma_{s_2, \dots, s_L, 0}|$. This is precisely what is accomplished by Lemmas 4 and 7. \square

3. The lower bound

The lower bound in Theorem 2 is an immediate consequence of the following estimate.

Lemma 8. *Let*

$$s_3 = \left\lfloor \frac{n^2(\sqrt{3}-1)}{6} \right\rfloor, \quad s_4 = \left\lfloor \frac{n^2(2-\sqrt{3})}{12} \right\rfloor.$$

For $c > 0$, let Γ_c be the collection of clique-decompositions of K_n in which there are $s_3 - \lfloor n^{2-c} \rfloor$ cliques with 3 vertices, s_4 cliques with 4 vertices, and the rest are cliques with 2 vertices. If $c > 0$ is sufficiently small then

$$|\Gamma_c| \geq \left(\frac{1+\sqrt{3}}{2} e^{\sqrt{3}/2-3} n - o(n) \right)^{n^{2/6}}.$$

To prove Lemma 8 we need a notion of ‘‘typicality’’ (called ‘‘quasirandomness’’ in [10]), closely related to the property in Lemma 5.

Definition 9. For an n -vertex, m -edge graph G , we define its density $p(G) = m/\binom{n}{2}$. We say that G is (ϵ, h) -typical if for every set A of at most h vertices of G , the vertices in A have $(1 \pm \epsilon)p(G)^{|A|}n$ common neighbours.

Note that if an n -vertex graph G with density p is (ϵ, h) -typical then it has

$$(1 \pm O_k(\epsilon)) p^{\binom{k}{2}} n^k / k!$$

k -cliques, for any $k \leq h$ (as for Equation (1), we count cliques vertex-by-vertex; this calculation also appears explicitly in [10, Proposition 2.8]).

Proof. Given a graph G and an integer k , we define its K_k -removal process as follows. Starting from the graph G , at each step we consider the set of all copies of K_k in our graph, choose one uniformly at random, and remove its edges. (Eventually we will run out of copies of K_k , at which point the process aborts).

We will need the following facts about the behaviour of the K_3 -removal process and the K_4 -removal process.

- (1) There is $a > 0$ such that the following holds. If we run the K_4 -removal process on K_n , then with probability $1 - o(1)$:
 - (a) the process does not abort before s_4 steps, and
 - (b) for each $t \leq s_4$, the graph at step t is $(n^{-a}, 3)$ -typical.
- (2) For every $a > 0$ there is $c > 0$ such that the following holds. Let G be an n -vertex graph with $m := \binom{n}{2} - 6s_4 = (3 - O(1/n))s_3$ edges which is $(n^{-a}, 2)$ -typical. If we run the K_3 -removal process on G , then with probability $1 - o(1)$:
 - (a) the process does not abort before $s_3 - n^{2-c}$ steps, and
 - (b) for each $t \leq s_3 - n^{2-c}$, the graph at step t is $(n^{-c}, 2)$ -typical.

For a simple proof of Fact (2), see [10, Theorem 4.1]. Fact (1) can be proved in basically exactly the same way (in fact it is slightly simpler, because we start from the complete graph instead of a general typical graph). See [7, Section 6] for some discussion of (a generalisation of) the K_k -removal process starting from a complete graph, which implies the desired result.

Now, we simply concatenate the K_4 -removal process and the K_3 -removal process. Indeed, starting from the complete graph K_n , we first run s_4 steps of the K_4 -removal process, then $s_3 - n^{2-c}$ steps of the K_3 removal process. In this way, either we abort or we produce a clique decomposition in Γ_c , in which our set of 4-cliques and our set of 3-cliques are both equipped with an ordering. Let Q be the set of outcomes of our concatenated process for which in each of the first s_4 steps, our graph is $(n^{-b}, 3)$ -typical, and in each of the next $s_3 - n^{2-c}$ steps, our graph is $(n^{-c}, 2)$ -typical.

The probability of each outcome in Q is at most

$$\prod_{t=1}^{s_4} \left((1 - n^{-\Omega(1)}) \left(\frac{\binom{n}{2} - 6t}{\binom{n}{2}} \right)^6 \frac{n^4}{24} \right)^{-1} \prod_{t=1}^{s_3 - n^{2-c}} \left((1 - n^{-\Omega(1)}) \left(\frac{\binom{n}{2} - 6s_4 - 3t}{\binom{n}{2}} \right)^3 \frac{n^3}{6} \right)^{-1}, \tag{4}$$

and by (1)–(2) above, these probabilities sum up to $1 - o(1)$. So, the number of outcomes in Q is at least $1 - o(1)$ divided by the expression in Equation (4). It follows that

$$|\Gamma_c| \geq \left(\frac{n^4}{24} \right)^{s_4} \left(\frac{n^3}{6} \right)^{s_3} \exp \left(6 \sum_{t=1}^{s_4} \log \left(\frac{\binom{n}{2} - 6t}{\binom{n}{2}} \right) + 3 \sum_{t=1}^{s_3} \log \left(\frac{\binom{n}{2} - 6s_4 - 3t}{\binom{n}{2}} \right) - n^{2-\Omega(1)} \right) / (s_4! s_3!).$$

(The difference between taking a sum up to s_3 and up to $s_3 - n^{2-c}$ is easily seen to contribute to the negligible $\exp(-n^{2-\Omega(1)})$ factor.)

Let $a_3 = 3s_3 / \binom{n}{2}$ and $a_4 = 6s_4 / \binom{n}{2}$, and note that $a_3 + a_4 = 1 - O(1/n)$. By Stirling’s approximation we compute that $\log |\Gamma_c|$ is at least

$$\begin{aligned} & \sum_{k=3}^4 s_k \left(\log \binom{n}{k} - \log s_k + 1 \right) + 6 \sum_{t=1}^{s_4} \log \left(\frac{\binom{n}{2} - 6t}{\binom{n}{2}} \right) + 3 \sum_{t=1}^{s_3} \log \left(\frac{\binom{n}{2} - 6s_4 - 3t}{\binom{n}{2}} \right) - n^{2-\Omega(1)} \\ &= \sum_{k=3}^4 s_k \left(\log \binom{n}{k} - \log s_k + 1 \right) \\ & \quad + 6 \binom{n}{2} \int_0^{a_4/6} \log(1 - 6t) dt + 3 \binom{n}{2} \int_0^{a_3/3} \log(1 - a_4 - 3t) dt - n^{2-\Omega(1)} \\ &= \sum_{k=3}^4 s_k \left(\log \binom{n}{k} - \log s_k + 1 \right) + \binom{n}{2} \int_0^{a_3+a_4} \log(1 - t) dt - n^{2-\Omega(1)} \\ &= \sum_{k=3}^4 s_k \left(\log \binom{n}{k} - \log s_k + 1 \right) - \binom{n}{2} - n^{2-\Omega(1)}. \end{aligned}$$

Substituting the values of s_3 and s_4 and simplifying (or alternatively, comparing with the expressions in the proof of Lemma 7) yields the desired result. □

Remark 10. In Lemma 8 we consider clique-decompositions that have a small number of “trivial” cliques with two vertices. We believe that it is possible to adapt the proof to avoid such cliques, but this requires some of Keevash’s deepest results on clique-decompositions of quasirandom graphs. Namely, for any constant k , Keevash’s machinery [7] allows one to estimate the number of K_k -decompositions of any dense quasirandom graph satisfying certain divisibility conditions (the number of edges should be divisible by $\binom{k}{2}$ and every degree should be divisible by $k - 1$; say such a graph is K_k -divisible). So, in order to prove a version of Lemma 8 in which no clique has exactly two vertices (thereby proving a version of Theorem 2 for proper linear spaces), it suffices to prove a suitable lower bound on the number of ways to partition the edges of K_n into a K_3 -divisible quasirandom graph with density $(\sqrt{3} - 1)/2 + o(1)$, a K_4 -divisible quasirandom graph with density $(2 - \sqrt{3})/2 + o(1)$ and a tiny “remainder graph” with $O(1)$ edges, itself decomposable into cliques which have more than two vertices. A suitable lower bound on the number of such graph partitions can be proved with some elementary number theory and the machinery of McKay and Wormald [12] for enumerating graphs with a given dense degree sequence (the remainder graph is just to handle divisibility issues, and it turns out we can always choose it to be either a copy of K_5 , a copy of K_7 , or a vertex-disjoint union $K_5 \cup K_7$).

References

- [1] D. M. Acketa, “On the enumeration of matroids of rank 2”, *Zb. Rad., Prirod.-Mat. Fak., Univ. Novom Sadu, Ser. Mat.* **8** (1978), p. 83-90.
- [2] N. Bansal, R. A. Pendavingh, J. G. van der Pol, “On the number of matroids”, *Combinatorica* **35** (2015), no. 3, p. 253-277.
- [3] L. M. Batten, A. Beutelspacher, *The theory of finite linear spaces*, Cambridge University Press, 1993.
- [4] W. M. B. Dukes, “On the number of matroids on a finite set”, *Sémin. Lothar. Comb.* **51** (2004), article no. B51g.
- [5] R. van der Hofstad, R. Pendavingh, J. G. van der Pol, “The number of partial Steiner systems and d-partitions”, *Adv. Comb.* **2022** (2022).
- [6] S. Janson, T. Łuczak, A. Ruciński, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley, 2000.
- [7] P. Keevash, “Counting designs”, *J. Eur. Math. Soc. (JEMS)* **20** (2018), no. 4, p. 903-927.
- [8] ———, “The existence of designs”, <https://arxiv.org/abs/1401.3665>, 2019.
- [9] D. E. Knuth, “The asymptotic number of geometries”, *J. Combinatorial Theory Ser. A* **16** (1974), p. 398-400.
- [10] M. Kwan, “Almost all Steiner triple systems have perfect matchings”, *Proc. Lond. Math. Soc.* **121** (2020), no. 6, p. 1468-1495.
- [11] N. Linial, Z. Luria, “An upper bound on the number of Steiner triple systems”, *Random Structures Algorithms* **43** (2013), no. 4, p. 399-406.
- [12] B. D. McKay, N. C. Wormald, “Wormald, Asymptotic enumeration by degree sequence of graphs of high degree”, *European J. Combin.* **11** (1990), no. 6, p. 565-580.
- [13] OEIS WikiContributors, “Index to OEIS: Matroids, sequences related to”, 2022, http://oeis.org/wiki/Index_to_OEIS:_Section_Mat#matroid.
- [14] J. Oxley, *Matroid theory*, second ed., Oxford Graduate Texts in Mathematics, vol. 21, Oxford University Press, Oxford, 2011.
- [15] R. A. Pendavingh, J. G. van der Pol, “Enumerating matroids of fixed rank”, *Electron. J. Combin.* **24** (2017), no. 1, article no. 1.8.
- [16] M. J. Piff, “An upper bound for the number of matroids”, *J. Combinatorial Theory Ser. B* **14** (1973), p. 241-245.
- [17] M. J. Piff, D. J. A. Welsh, “The number of combinatorial geometries”, *Bull. London Math. Soc.* **3** (1971), p. 55-56.
- [18] J. G. van der Pol, “Large matroids: enumeration and typical properties”, PhD Thesis, Mathematics and Computer Science, Technische Universiteit Eindhoven, Netherlands, 2017.
- [19] E. E. Shult, *Points and lines. Characterizing the classical geometries*, Universitext, Springer, 2011.
- [20] D. J. A. Welsh, *Matroid theory*, London Mathematical Society Monographs, vol. 8, Academic Press [Harcourt Brace Jovanovich, Publishers], 1976.