

On the rank of a random binary matrix

Colin Cooper*

Department of Computer Science
Kings College
London WC2R 2LS
U.K.

cooper.colin@gmail.com

Alan Frieze†

Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh PA15213
U.S.A.

alan@random.math.cmu.edu

wes@math.cmu.edu

Wesley Pegden‡

Submitted: Aug 14, 2018; Accepted: Sep 24, 2019; Published: Oct 11, 2019

© The authors. Released under the CC BY license (International 4.0).

Abstract

We study the rank of a random $n \times m$ matrix $\mathbf{A}_{n,m;k}$ with entries from $GF(2)$, and exactly k unit entries in each column, the other entries being zero. The columns are chosen independently and uniformly at random from the set of all $\binom{n}{k}$ such columns.

We obtain an asymptotically correct estimate for the rank as a function of the number of columns m in terms of c, n, k , and where $m = cn/k$. The matrix $\mathbf{A}_{n,m;k}$ forms the vertex-edge incidence matrix of a k -uniform random hypergraph H . The rank of $\mathbf{A}_{n,m;k}$ can be expressed as follows. Let $|C_2|$ be the number of vertices of the 2-core of H , and $|E(C_2)|$ the number of edges. Let m^* be the value of m for which $|C_2| = |E(C_2)|$. Then w.h.p. for $m < m^*$ the rank of $\mathbf{A}_{n,m;k}$ is asymptotic to m , and for $m \geq m^*$ the rank is asymptotic to $m - |E(C_2)| + |C_2|$.

In addition, assign i.i.d. $U[0, 1]$ weights $X_i, i \in \{1, 2, \dots, m\}$ to the columns, and define the weight of a set of columns S as $X(S) = \sum_{j \in S} X_j$. Define a basis as a set of $n - \mathbf{1}(k \text{ even})$ linearly independent columns. We obtain an asymptotically correct estimate for the minimum weight basis. This generalises the well-known result of Frieze [On the value of a random minimum spanning tree problem, *Discrete Applied Mathematics*, (1985)] that, for $k = 2$, the expected length of a minimum weight spanning tree tends to $\zeta(3) \sim 1.202$.

Mathematics Subject Classifications: 60C05, 15B52

*Research supported in part by EPSRC grant EP/M005038/1

†Research supported in part by NSF Grant DMS1661063

‡Research supported in part by NSF grant DMS1363136

1 Introduction

Let $\Omega_{n,k}$ denote the set of vectors of length n , with 0, 1 entries, with exactly k 1's, all other entries being zero. The addition of entries is over the field GF_2 , i.e., the vector addition is over $(GF_2)^n$. Let $\mathbf{A}_{n,m;k}$ be the random $n \times m$ matrix where the columns form a random m -subset of $\Omega_{n,k}$.

In a recent paper [7], we studied the binary matroid $\mathcal{M}_{n,m;k}$ induced by the columns of $\mathbf{A}_{n,m;k}$. It was shown that for any fixed binary matroid M , there were constants k_M, L_M such that if $k \geq k_M$ and $m \geq L_M n$ then w.h.p. $\mathcal{M}_{n,m;k}$ contains M as a minor. The paper [7] contributes to the theory of random matroids as developed by [1], [3], [11], [13], [14]. In this paper we study a related aspect of $\mathbf{A}_{n,m;k}$, namely its rank, and improve on results from Cooper [5]. As a consequence of the precise estimate of rank in Theorem 1 we can give an expression, (5), for the solution value of the following optimization problem.

Suppose that we assign i.i.d. $U[0, 1]$ weights $X_{\mathbf{c}}$ to the vectors $\mathbf{c} \in \Omega_{n,k}$ and let the weight of a set of columns S be $X(S) = \sum_{\mathbf{c} \in S} X_{\mathbf{c}}$. Define a *basis* as a set of $n - 1$ (k even) linearly independent columns. What is the expected weight $W_{n,k}$ of a minimum weight basis? When $k = 2$ this amounts to estimating the expected length of a minimum weight spanning tree of K_n which has the limiting value of $\zeta(3)$, see Frieze [8].

Our result on the rank of $\mathbf{A}_{n,m;k}$ takes a little setting up. Let $H = H_{n,m;k}$ denote the random k -uniform hypergraph with vertex set $[n]$ and m random edges taken from $\binom{[n]}{k}$. There is a natural bijection between $\mathbf{A}_{n,m;k}$ and $H_{n,m;k}$ in which column \mathbf{c} is replaced by the set $\{i : \mathbf{c}_i = 1\}$. The ρ -core of a hypergraph H (if it is non-empty) is the maximal set of vertices that induces a sub-hypergraph of minimum degree ρ . The 2-core $C_2 = C_2(H)$ plays an important role in our first theorem.

1.1 Matrix Rank

Notation: We write $X_n \approx Y_n$ for sequences $X_n, Y_n, n \geq 0$ if $X_n = (1 + o(1))Y_n$ as $n \rightarrow \infty$. Our results are asymptotic in $n, m(n)$, as $n, m \rightarrow \infty$, whereas k is a fixed positive integer.

We will use some results on the 2-core of random hypergraphs. The size of the 2-core has been asymptotically determined, see for example Cooper [6] or Molloy [12]; we recall the basic w.h.p. results here. In random graphs $G_{n,m} = H_{n,m;2}$ the 2-core grows gradually with m following the emergence of the first cycle of size $O(\log n)$. For $k \geq 3$, the 2-core is either empty or of linear size and emerges around some threshold value \hat{m}_k . Initially above \hat{m}_k the 2-core has more vertices than edges, and there is a larger value m^* , around which the number of vertices and edges becomes the same. Below m^* the rank of the 2-core grows asymptotically as the number of edges, and above m^* as the number of vertices.

To describe the size of the 2-core, we parameterise m as $m = cn/k$, $c = O(1)$ and consider the equation

$$x = (1 - e^{-cx})^{k-1}. \tag{1}$$

For $k \geq 3$, define \widehat{c}_k by

$$\widehat{c}_k = \min \left\{ c : x = (1 - e^{-cx})^{k-1} \text{ has a solution } x_c \in (0, 1] \right\}.$$

It is known that $c < \widehat{c}_k$ implies that $C_2 = \emptyset$. If $c > \widehat{c}_k$, $c = O(\log n)$, let x_c be the largest solution to (1) in $[0, 1]$. Then q.s.¹

$$\left| |C_2| - n(x_c^{1/(k-1)} - cx_c + cx_c^{k/(k-1)}) \right| \leq n^{3/4}, \quad (2)$$

$$\left| |E(C_2)| - n(cx_c^{k/(k-1)}/k) \right| \leq n^{3/4}. \quad (3)$$

We note for future reference that using (1), the term $x^{1/(k-1)} - cx + cx^{k/(k-1)}$ in (2) can be written as $1 - e^{-cx}(1 + cx)$.

Let c_k^* be the value of c for which the 2-core has asymptotically the same number of vertices and edges. More precisely, we use (2) and (3) to define c_k^* by

$$c_k^* := \min \left\{ c \geq \widehat{c}_k : x_c^{1/(k-1)} - cx_c + cx_c^{k/(k-1)} = \frac{cx_c^{k/(k-1)}}{k} \right\}. \quad (4)$$

Define m_k^* by $m_k^* = c_k^* n/k$. We will prove,

Theorem 1. *If $m = O(n)$ then w.h.p.*

$$\text{rank}(\mathbf{A}_{n,m;k}) \approx \begin{cases} |E(H)| & m < m_k^* \\ |E(H)| - |E(C_2)| + |C_2| & m \geq m_k^* \end{cases}$$

Note that when $k = 2$ we have $c_2^* = 0$ and the theorem follows from the fact that an isolated tree with t edges induces a sub-matrix of rank t in $\mathbf{A}_{n,m;k}$. We therefore concentrate on the case $k \geq 3$.

Using (2) and (3), we can express Theorem 1 directly in terms of c by

Corollary 2. *Suppose that $k \geq 3$ and $m = cn/k$. Then, w.h.p.*

$$\text{rank}(\mathbf{A}_{n,m;k}) \approx \begin{cases} m & c < c_k^* \\ m - mx_c^{k/(k-1)} + n(x_c^{1/(k-1)} - cx_c + cx_c^{k/(k-1)}) & c \geq c_k^* \end{cases} \quad (5)$$

Around $m = n(\log n + d_n)/k$, where $d_n = o(\log n)$, the remaining vertices of degree one in H disappear, and $\mathbf{A}_{n,m;k}$ has full rank up to parity, i.e., $\text{rank}(\mathbf{A}_{n,m;k}) = n^*$ where

$$n^* = n - \mathbf{1}(k \text{ even}).$$

Theorem 3. *Suppose that $k \geq 3$.*

(i) *Given a constant $A > 0$, there exists $\gamma = \gamma(A)$ such that for $m \geq \gamma n \log n$,*

$$\Pr(\text{rank}(\mathbf{A}_{n,m;k}) < n^*) = o(n^{-A}).$$

¹A sequence \mathcal{E}_n of events occurs *quite surely* (q.s.) if $\Pr(-\mathcal{E}_n) = O(n^{-C})$ for any constant $C > 0$.

(ii) If $m = n(\log n + d_n)/k$ then

$$\lim_{n \rightarrow \infty} \Pr(\text{rank}(\mathbf{A}_{n,m;k} = n^*)) = \begin{cases} 0 & d_n \rightarrow -\infty \\ e^{-e^{-d}} & d_n \rightarrow d \\ 1 & d_n \rightarrow +\infty. \end{cases}$$

We can easily modify the proof of part (ii) of Theorem 3 to give the following hitting time version. Suppose that we randomly order the columns of $\mathbf{A}_{n,M;k}$ where $M = \binom{n}{k}$. Let \mathbf{M}_m denote the matrix defined by the first m columns in this order.

$$m_1 = \min \{m : \mathbf{M}_m \text{ has } n^* \text{ non-zero rows}\} \text{ and let } m^* = \min \{m : \mathbf{M}_m \text{ has rank } n^*\}.$$

Theorem 4. $m_1 = m^*$ w.h.p.

Some time after completion of this manuscript, we learnt from Amin Coja-Oghlan of an independent proof of Theorem 1, see [2].

1.2 Minimum Weight Basis

The expression (5) enables us to estimate the expected optimal value to the minimum weight basis problem defined above. Suppose that we assign i.i.d. $U[0, 1]$ weights $X_{\mathbf{c}}$, $\mathbf{c} \in \Omega_{n,k}$ to the $|\Omega_{n,k}| = \binom{n}{k}$ distinct vectors with exactly k unit entries, all other entries being zeroes. The weight of a set of columns C is $X(C) = \sum_{\mathbf{c} \in C} X_{\mathbf{c}}$. Let $W_{n,k}$ be the minimum weight of any basis of $n^* = n - 1$ (k even) linearly independent columns, chosen from the $\binom{n}{k}$ column vectors $\mathbf{c} \in \Omega_{n,k}$. Define the random matrix $\mathbf{A}_{n,p;k}$ to consist of the vectors $\mathbf{c} \in \Omega_{n,k}$ with weight $X_{\mathbf{c}}$ at most p .

We show in Section 3 below that if $W_{n,k}$ denotes the weight of a minimum weight basis then

$$\mathbf{E}(W_{n,k}) = \int_{p=0}^1 (n^* - \mathbf{E}(\text{rank}(\mathbf{A}_{n,p;k}))) dp. \quad (6)$$

Corollary 2 and Theorem 3 can be substituted into (6) to yield an asymptotic formula for $W_{m,k}$.

Theorem 5. Let $x = x(c)$ be the largest solution of $x = (1 - e^{-cx})^{k-1}$ in $(0, 1]$, then

$$\frac{n^{k-2}}{(k-1)!} \mathbf{E}(W_{n,k}) \approx c_k^* \left(1 - \frac{c_k^*}{2k}\right) + \int_{c_k^*}^{\infty} \left(e^{-cx} \left(1 + \frac{(k-1)cx}{k}\right) - \frac{c}{k}(1-x) \right) dc \quad (7)$$

We note the remarkable fact that, by the result of Frieze [8], for $k = 2$ and with $c_2^* = 0$, the expression in (7) must equal $\zeta(3)$. We have numerically estimated the first few values as a function of k :

k	2	3	4	5	6	7	8	9	10
$\frac{n^{k-2}}{(k-1)!} \mathbf{E}(W_{n,k})$	$\zeta(3) \approx 1.202$	1.563	2.021	2.507	3.003	3.501	4.000	4.500	5.000

It appears the values are getting close to $k/2$ as k grows, and this is indeed the case.

Theorem 6. For $k \geq 3$, and some ε_k , $|\varepsilon_k| \leq 5$,

$$\lim_{n \rightarrow \infty} \frac{n^{k-2}}{(k-1)!} \mathbf{E}(W_{n,k}) = \frac{k}{2} (1 + \varepsilon_k e^{-k}). \quad (8)$$

2 Matrix Rank

We study the random matrix \mathbf{A}_m distributed as $\mathbf{A}_{n,m;k}$, with corresponding hypergraph H_m distributed as $H_{n,m;k}$. We let $c = km/n$.

The first step of our proof is to “peel off” edges of the hypergraph H_m , and thus columns of the matrix \mathbf{A}_m , containing vertices of degree 1.

In particular, we set $H_m := H_m$, and then, recursively, so long as H_i contains a vertex x_i of degree 1, then for the edge $e_i \ni x_i$ in H_i , we set

$$\begin{aligned} E(H_{i-1}) &= E(H_i) \setminus \{e_i\} \\ V(H_{i-1}) &= V(H_i) \setminus \{x \in e_i \mid \deg_{H_i}(x) = 1\}. \end{aligned}$$

In a corresponding sequence $\{\mathbf{A}_i\}$ beginning from \mathbf{A}_m , we obtain \mathbf{A}_{i-1} from \mathbf{A}_i by removing the column c_i corresponding to e_i , and the (at least one) rows whose only 1s were in that column. Note that for all $i < m$ for which \mathbf{A}_i is defined, we have

$$\text{rank}(\mathbf{A}_i) = \text{rank}(\mathbf{A}_{i+1}) - 1.$$

This recursion terminates at

$$\mathbf{C}_2 = \mathbf{A}_{m_2}, \quad (9)$$

where $m_2 = m - m_1$ is the number of edges in the the 2-core of the hypergraph H , and moreover, we have that H_{m_2} is precisely the 2-core of H . Thus we have that

$$\text{rank}(\mathbf{A}_m) = m_1 + \text{rank}(\mathbf{C}_2). \quad (10)$$

We consider the cases which control the behavior of the rank of the 2-core $\mathbf{C}_2 = H_{m_2}$ of H . We use a theorem of Pittel and Sorkin [15] which we state here for completeness. A system of $M \times N$ equations is uniformly constrained if each variable N appears at least twice. The theorem of [15] as reproduced below describes the transpose of our formulation, i.e., A is an $M \times N$ matrix, and thus full row rank of A corresponds to full column rank of the 2-core matrix.

Theorem ([15] Theorem 2). Let $Ax = b$ be a uniformly random constrained k -XORSAT instance with M equations and N variables, with $k \geq 3$ and $N, M \rightarrow \infty$ with $\liminf M/N > 2/k$. Then, for any $\omega(N) \rightarrow \infty$, if $M = N - \omega(N)$ then $Ax = b$ is almost surely satisfiable, with satisfiability probability $1 - O(M^{-(k-2)} + \exp(-0.59\omega(N)))$, while if $M = N + \omega(N)$ then $Ax = b$ is almost surely unsatisfiable, with satisfiability probability $O(2^{-\omega(N)})$.

Let $N = |V(C_2)|$ and $M = m_2 = |E(C_2)|$ be the number of rows and columns of the 2-core matrix \mathbf{C}_2 . The columns associated with the 2-core \mathbf{C}_2 are distributed as uniformly random, subject to each vertex/row of the 2-core being in at least two columns.

Case 1: $c < c_k^*$.

Here $M < N$. It follows from the above Theorem of Pittel and Sorkin [15], that the rank of the columns $\mathbf{c}_{m_1+1}, \mathbf{c}_{m_1+2}, \dots, \mathbf{c}_m$ is $\approx M = m_2 = m - m_1$.

For this case the first claim of (5), and Theorem 1, have been verified.

Case 2: $c \geq c_k^*$.

Here $M > N$. To prove Theorem 1 for $c \geq c_k^*$ we need to verify that w.h.p.

$$\text{rank}(\mathbf{C}_2) \approx |V(C_2)|. \quad (11)$$

In this case we need some basic facts about hypergraphs. We say a hypergraph H is *linear* if edges only intersect in at most one vertex. We define a k -uniform *cactus* as follows. A single edge is a cactus. An $(\ell + 1)$ -edge cactus C' is the structure obtained from an ℓ -edge cactus C with vertex set $V(C)$, $|V(C)| = (k - 1)\ell + 1$ as follows. Choose $x \in V(C)$ and let $V(C') = V(C) \cup \{v_1, \dots, v_{k-1}\}$ where $\{v_1, \dots, v_{k-1}\}$ is disjoint from $V(C)$. The edge set $E(C')$ of C' is $E(C) \cup \{e'\}$ where $e' = \{x, v_1, \dots, v_{k-1}\}$. We need the following simple lemma.

Lemma 7. *A connected k -uniform simple hypergraph C with no cycles is a cactus.*

Proof. This can easily be verified by induction. We simply remove one terminal edge $e = \{v_1, v_2, \dots, v_k\}$ of a longest path P . We can assume here that v_2, \dots, v_k are all of degree one, else P can be extended. Deleting e gives a new connected hypergraph C' which is a cactus by induction. \square

For a k -uniform linear hypergraph H let $L(H) = (k - 1)|E(H)| + 1$.

Lemma 8. *Let H be a connected k -uniform linear hypergraph.*

(a) $|V(H)| \leq L(H)$.

(b) $|V(H)| = L(H)$ if and only if H does not contain any cycles.

(c) *By deleting at most $L(H) - |V(H)|$ edges we can create a subgraph H' with $V(H') = V(H)$ and no cycles.*

Proof. We consider two cases:

Case 1: H contains no cycles.

In this case, we consider a longest path of edges in H ; that is consider a longest sequence e_1, e_2, \dots, e_ℓ such that for each $1 < i < \ell$, e_i intersects e_{i-1} , e_{i+1} , and no other edges in the sequence. Since the path is longest and H has no cycles, we know that e_ℓ intersects no edge in H other than $e_{\ell-1}$.

In particular, we define a hypergraph H' with $E(H') = E(H) \setminus \{e_\ell\}$ and $V(H') = V(H) \setminus (e_\ell \setminus e_{\ell-1})$. H' has one fewer edge and $k - 1$ fewer vertices than H , so we have $L(H) = |V(H)|$ by induction, proving the Lemma for this case.

Case 2: H contains a cycle C .

In this case, we consider an edge e in a cycle C of H . Removing the edge e leaves a hypergraph on the same vertex set with one fewer edge and with at most $k - 1$ connected components (counting isolated vertices as connected components). Applying the Lemma inductively to each component, we see that the sum of $L(H_i)$ over the $(k - 1)$ components H_i of $H \setminus e$ satisfies

$$\sum_{i=1}^{k-1} L(H_i) \leq L(H) - (k - 1) + (k - 2) \leq L(H) - 1,$$

since removing e decreases the sum by $k - 1$, while the additive term in the definition of $L(H)$ inflates the sum by at most $(k - 2)$ (as the number of components has increased by up to $k - 2$). On the other hand we of course have

$$\sum_{i=1}^{k-1} |V(H_i)| = |V(H)|.$$

We now apply parts (a) and (c) of the Lemma to each component by induction, and conclude that the Lemma does hold for H . \square

In the following lemma we prove a property of $H_{n,m;k}$. It will be more convenient to work with $H_{n,p;k}$ where $m = \binom{n}{k}p$. We use the fact that for any hypergraph property \mathcal{H} that is monotone increasing or decreasing with respect to adding edges,

$$\Pr(H_{n,m;k} \in \mathcal{H}) \leq O(1) \Pr(H_{n,p;k} \in \mathcal{H}). \quad (12)$$

This is well-known for graphs and is essentially a property of the binomial random variable, $E(H_{n,p;k})$, the number of edges of $H_{n,p;k}$.

Similarly, if \mathcal{A} is a matrix property that is monotone increasing or decreasing with respect to adding columns, then

$$\Pr(\mathbf{A}_{n,m;k} \in \mathcal{A}) \leq O(1) \Pr(\mathbf{A}_{n,p;k} \in \mathcal{A}). \quad (13)$$

Lemma 9. *Suppose that $m = O(n \log n)$.*

(a) *Let $\alpha < 1$ be a positive constant. With probability $1 - o(n^{-1})$, for every set of vertices S of size $\ell_0 = \log^{1/2} n \leq s \leq s_0 = n^{1-\alpha}$ we have that $L(S) \leq s + \lfloor \theta s \rfloor$, where $\theta = \frac{1}{\log^{1/4} n}$. Here $H[S]$ is the hypergraph of edges belonging completely to S .*

(b) *Then w.h.p., there are at most $n^{o(1)}$ vertices in cycles of size at most $\log^{1/2} n$.*

Proof. (a) We can use (12) here with $p = \frac{C \log n}{n^{k-1}}$ for some $C = O(1)$ satisfying $m = \binom{n}{k}p$. Let $s_1 = s + \lfloor \theta s \rfloor + 1$. The expected number of sets failing this property can be bounded by

$$\sum_{s=\ell_0}^{s_0} \binom{n}{s} \sum_{L \geq s_1} \binom{s}{L/(k-1)} \left(\frac{C \log n}{n^{k-1}} \right)^{L/(k-1)}$$

$$\begin{aligned}
&\leq \sum_{s=\ell_0}^{s_0} \left(\frac{ne}{s}\right)^s \sum_{L \geq s_1} \left(\frac{Ces^k \log n(k-1)}{k!Ln^{k-1}}\right)^{L/(k-1)} \\
&\leq \sum_{s=\ell_0}^{s_0} \sum_{L \geq s_1} (Ce^2 \log n)^L \left(\frac{s}{n}\right)^{L-s} \left(\frac{s}{L}\right)^{L/(k-1)} \\
&\leq \sum_{s=\ell_0}^{s_0} \sum_{L \geq s_1} \left((Ce^2 \log n) \left(\frac{s}{n}\right)^{1-s/L}\right)^L
\end{aligned} \tag{14}$$

Let $u_{s,L}$ denote the summand in (14). Then we have

$$\begin{aligned}
u_{L,s} &\leq \left((Ce^3 \log n)^{2\alpha^{-1}} \left(\frac{s}{n}\right)^\theta\right)^s && L \leq 2\alpha^{-1}s. \\
u_{L,s} &\leq \left((Ce^3 \log n) \left(\frac{s}{n}\right)^{1-\alpha/2}\right)^L && L > 2\alpha^{-1}s.
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{s \geq \ell_0} \sum_{L \geq s_1} u_{s,L} &\leq \sum_{s=\ell_0}^{s_0} \sum_{L=s+\lceil \theta s \rceil}^{2\alpha^{-1}s} n^{-(\alpha-o(1))\theta s} + \sum_{s=\ell_0}^{s_0} \sum_{L \geq 2\alpha^{-1}s} n^{-(1-o(1))\alpha L/2} \\
&\leq 2\alpha^{-1}s_0 \sum_{s=\ell_0}^{s_0} n^{-(\alpha-o(1))\theta s} + \sum_{s=\ell_0}^{s_0} n^{-(1-o(1))s/2} \\
&= o(n^{-1}).
\end{aligned} \tag{15}$$

(b) The expected number of vertices in small cycles can be bounded by

$$\sum_{\ell=2}^{\log^{1/2} n} \binom{n}{(k-1)\ell} ((k-1)\ell)! p^\ell \leq \sum_{\ell=2}^{\log^{1/2} n} (n^{k-1} p)^\ell \leq \sum_{\ell=2}^{\log^{1/2} n} (C \log n)^\ell = n^{o(1)}.$$

Part (b) now follows from the Markov inequality. \square

2.1 Growth of the mantle

We now consider the change in the rank of the sub-matrix \mathbf{C}_2 of the edge-vertex incidence matrix \mathbf{A}_m (see (9)) corresponding to the 2-core of the column hypergraph, caused by adding a column to \mathbf{A}_m . In this section, we will assume in our calculations that no two edges share more than one vertex, and that the 2-core consists of a single connected component. This does not affect our asymptotic analysis because simple first-moment calculations show that:

1. There are only a bounded number of edges sharing more than one vertex, and

2. Any subset of the random hypergraph of minimum degree at least 2 must be of linear size; together with (16), below, this then implies that the 2-core can only have one connected component in the present regime, since the appearance of another component at any state would increase the size of the 2-core by too much.

So suppose now that the addition of e increases the size of the 2-core. Let A denote the set of additional vertices and F denote the set of additional edges added to C_2 by the addition of e , where $A \subset V(F)$. We include e in F .

We remark first that with c, x as in (1), then (2) and (3) state that q.s.

$$||C_2| - n(x_c^{1/(k-1)} - cx_c + cx_c^{k/(k-1)})| \leq n^{3/4}, \quad \text{and} \quad ||E(C_2)| - mx^{k/(k-1)}| \leq n^{3/4}. \quad (16)$$

Therefore we can assume that adding an edge to \mathbf{A}_m can only increase C_2 , $E(C_2)$ by at most $O(n^{3/4})$. We use Lemma 9 with $\alpha = 3/4$ in our discussion of the hypergraph F .

Obviously the increase in rank from adding F to the 2-core is bounded above by the size of the vertex-set A . To bound it from below, we proceed as follows:

Case 1: First consider the case where there are no cycles in F . We will show that the rank increases by precisely the number of new vertices.

Let $|A| = k$. We will define an ordering a_1, \dots, a_k of A and a corresponding ordering f_1, \dots, f_k of a subset of F . To begin, we claim there must exist $v \in A$ and $v \in f \in F$, $f \neq e$, such that $f \setminus \{v\} \subseteq C_2$. For this consider a longest path e_1, \dots, e_ℓ of edges in F . Since the hypergraph is simple and contains no cycles, we have that $e_\ell \cap (\bigcup_{i=1}^{\ell-1} e_i) = e_\ell \cap e_{\ell-1} = \{v\}$ for some single vertex v . On the other hand, all vertices of e_ℓ must have degree 2 in $F \cup C_2$, and so $e_\ell \setminus v$ must lie entirely in C_2 . We set $f_1 = e_\ell$, $a_1 = v$, and then we remove f_1 from F and a_1 from A , defining $C_2^1 = C_2 \cup f_1$ (though it is not a two-core of any hypergraph), and apply induction to obtain the sequences a_1, \dots, a_k , f_1, \dots, f_k , and the corresponding sequence C_2^i defined by $C_2^0 = C_2$, and $C_2^{i+1} = C_2^i \cup f_{i+1}$.

These sequences have the property that

$$\text{rank}(C_2^{i+1}) = \text{rank}(C_2^i) + 1,$$

since the edge f_i added to C_2^i in step $i + 1$ contains exactly one vertex outside of C_2^i . (In the matrix, we are adding a column containing a 1 in a row which previously had no 1's).

In particular, the rank in this case increases by exactly the size of A .

Case 2: The total contribution to the rank of the 2-core in $m = O(n \log n)$ steps from the case where F contains a cycle of length at most $\log^{1/2} n$ can be bounded by $n^{3/4+o(1)}$. This follows from Lemma 9(b) and (16). This is negligible, since the core has size $\Omega(n)$ in the regime we are discussing.

Case 3: Suppose that F contains cycles of size at least $\log^{1/2} n$ which we remove by deleting s edges. When we do this we may lose up to ks vertices from A . Let the resulting vertex set be A' and edge set be F' . Up to ks vertices of A' may have degree 1. Attach these vertices to C_2 using disjoint edges to give edge set F'' . All vertices of A' now have degree at least 2 in F'' and F'' has no cycles. According to the argument in Case 1, the increase in rank due to adding F'' is $|A'| \geq |A| - ks$ and this is at most ks larger than

the increase in rank due to adding F' . Thus the increase in rank due to adding $F \supseteq F'$ is at least $|A| - 2ks$ and at most $|F| \leq |A| + s + 1$. It follows from Lemma 8(c) and Lemma 9(a) that $s = O(|A|/\log^{1/4} n)$.

In summary we find that if $m = O(n \log n)$ and $m \geq c^*n/k$ then, with probability $1 - o(n^{-1})$, the rank of \mathbf{C}_2 satisfies

$$\left(1 - O(1/\log^{1/4} n)\right) |C_2| \leq \text{rank}(\mathbf{C}_2) \leq |C_2|. \quad (17)$$

The upper bound follows because the rank of \mathbf{C}_2 is at most the number of rows in \mathbf{C}_2 . This proves (11). To finish the proof of Theorem 1 we require that (17) remains true if we take expectations. For this we use the error probability of $o(n^{-1})$ in (15).

2.2 Proof of Theorem 3

Proof of part (i):

Given a set of rows S of size $s = |S|$, the number of choices of column (distinct edges) that have an odd number of non-zero entries in S is

$$T_{s,k} = \binom{s}{1} \binom{n-s}{k-1} + \binom{s}{3} \binom{n-s}{k-3} + \cdots + \binom{s}{k}.$$

If $\text{rank}(\mathbf{A}_{n,p;k}) < n^*$ then there exists a set S of rows such that (i) each column of $\mathbf{A}_{n,p;k}$ has an even number of non-zero entries j in S and (ii) $|S| \leq n^*$. For a fixed S , denote this event by \mathcal{B}_S and note that it is monotone decreasing. Then

$$\Pr(\mathcal{B}_S) = (1-p)^{T_{s,k}}. \quad (18)$$

For $s \geq k$,

$$T_{s,k} \geq \binom{s}{1} \binom{n-s}{k-1} + \binom{s}{k} = \frac{s}{(k-1)!} \left(\frac{s^{k-1}}{k} + (n-s)^{k-1} \right) (1 + o(1))$$

The bracketed term on the right hand side is minimized when $s = \alpha n$ where $\alpha = k^{1/(k-2)} / (1 + k^{1/(k-2)})$. Let $\beta_k = (\alpha^{k-1}/k + (1-\alpha)^{k-1})$ then

$$T_{s,k} \geq \beta_k s \frac{n^{k-1}}{(k-1)!} (1 + o(1)).$$

We can choose $p = \frac{(A+2)\log n}{\beta_k \binom{n-1}{k-1}}$ and then use monotonicity of rank as a function of p to claim the result for larger p .

$$\begin{aligned} \Pr(\exists S : \mathcal{B}_S \text{ occurs}) &\leq \sum_{s=1}^{n^*} \binom{n}{s} (1-p)^{T_{s,k}} \\ &\leq \sum_{s=1}^{n^*} \left(\frac{ne}{s} \cdot \exp \left\{ -p \beta_k \frac{n^{k-1}}{(k-1)!} (1 + o(1)) \right\} \right)^s \end{aligned} \quad (19)$$

$$\leq \sum_{s=1}^{n^*} n^{-(A+1+o(1))s} = O\left(\frac{1}{n^{A+1+o(1)}}\right).$$

We use (13) to transfer this bound to $\mathbf{A}_{n,m;k}$.

Proof of part (ii).

Let $m = n(\log n + c_n)/k$. We first observe that if Z_s denotes the number of sets of $s = O(1)$ empty rows then

$$\begin{aligned} \mathbf{E}(Z_s) &= \binom{n}{s} \frac{\binom{\binom{n-s}{k}}{m}}{\binom{\binom{n-s}{k}}{m}} = \binom{n}{s} \prod_{i=0}^{m-1} \frac{\binom{n-s}{k} - i}{\binom{n-s}{k}} = \binom{n}{s} \left(\frac{\binom{n-s}{k}}{\binom{n-s}{k}}\right)^m \left(1 + O\left(\frac{m^2}{n^k}\right)\right) \\ &\approx \frac{n^s}{s!} \cdot \prod_{i=0}^{k-1} \left(1 - \frac{s}{n-i}\right)^m = \frac{n^s}{s!} \cdot \prod_{i=0}^{k-1} \exp\left\{-\frac{ms}{n} + O\left(\frac{m}{n^2}\right)\right\} \approx \frac{n^s}{s!} e^{-skm/n} \approx \frac{e^{-cs}}{s!}. \end{aligned} \quad (20)$$

Thus if $c_n \rightarrow \infty$, $\mathbf{E}(Z_1) \rightarrow 0$, and if $c_n \rightarrow -\infty$, $\mathbf{E}(Z_1) \rightarrow \infty$. Straightforward arguments complete Theorem 3(ii) for these cases.

Assume next that $c_n \rightarrow c$. The method of moments applied to (20) implies that Z_1 is asymptotically Poisson with mean e^{-c} and so

$$\Pr(Z_1 = 0) \approx e^{-e^{-c}}. \quad (21)$$

The final step is to prove (w.h.p) that when $p = (\log n + c_n)/\binom{n-1}{k-1}$, $c_n \rightarrow c$ constant, the only obstruction to $\text{rank}(\mathbf{A}_{n,p;k}) = n^*$ is the existence of empty rows ($Z_1 > 0$). As in part (i) above, going back to (19) with $p = (\log n + c_n)/\binom{n-1}{k-1}$ we see that we only need to consider $2 \leq s \leq e^{2+\beta_k|c_n|} n^{1-\beta_k}$. For these values of s , $T_{s,k}$ is bounded below by $s\binom{n-s}{k-1} \approx s\binom{n-1}{k-1}$. Similar to the derivation of (19), we see we can bound the probability of the event $\{\exists S : \mathcal{B}_S\}$ from above by

$$\Pr(\mathcal{B}_S) \leq \left(\frac{3n}{s} \cdot \exp\left\{-p\binom{n-1}{k-1}\right\}\right)^s = \left(\frac{O(e^{-c})}{s}\right)^s.$$

Thus, for $|c|$ constant, with $s_1 = e^{2+\beta_k|c_n|} n^{1-\beta_k}$

$$\Pr(\exists S, \log \log n \leq |S| \leq s_1 : \mathcal{B}_S \text{ occurs}) \leq \sum_{s=\log \log n}^{s_1} \left(\frac{O(e^{-c})}{s}\right)^s = o(1). \quad (22)$$

Finally we consider $2 \leq s \leq L = \log \log n$. Given a set S , the number of choices of column that have an odd number of non-zero entries in S (Type A columns) is given by $T_{s,k}$ above. The number of choices of columns that have an even number of non-zero entries in S (Type B columns) is

$$R_{s,k} = \binom{s}{2} \binom{n-s}{k-2} + \cdots + \binom{s}{k-1} (n-s).$$

For $s \leq L$, $R_{s,k} \leq s^2 n^{k-2}$. The expected number μ_s of sets S with no Type A columns and at least one Type B column is

$$\mu_s = \binom{n}{s} (1 - (1-p)^{R_{s,k}}) (1-p)^{T_{s,k}} \leq \frac{n^s}{s!} (pR_{s,k}) e^{-ps \binom{n-1}{k-1} (1+o(1))} = O\left(\frac{\log n}{n}\right) e^{-cs}.$$

Thus, for constant c ,

$$\sum_{s=2}^L \mu_s = o(1). \tag{23}$$

Thus w.h.p. there is no set of $2 \leq s \leq \log \log n$ rows where the dependency does not come from the rows all being zero. \square

2.3 Proof of Theorem 4

Because c in (21) is arbitrary and having a zero row is a monotone decreasing event, we can see that if $m_0 = n(\log n - \log \log n)/k$ then $Z_1 = Z_1(m_0) > 0$ w.h.p. The reader can easily check that equations (22) and (23) continue to hold. It follows that w.h.p. the rank of \mathbf{M}_{m_0} is $n^* - Z_1$. It then follows that $m_1 = m^*$ if we never add a column that reduces the number of non-zero rows by more than one. Now (21) implies that the expected number of zero rows in \mathbf{M}_{m_0} is $O(\log n)$ and so $Z_1 \leq \log^2 n$ w.h.p. So given this, the probability we add a column that reduces the number of non-zero rows by more than one in the next $O(n \log n)$ column additions, is $O(n \log n \times ((\log^2 n)/n)^2) = o(1)$.

3 Minimum Weight Basis

The first task here is to prove (6). Let $B_{n,k}$ denote a minimum weight basis and let $W_{n,k}$ denote its weight. For a given a real number X we can write

$$X = \int_{p=0}^X dp = \int_{p=0}^1 1_{p \leq X} dp.$$

Thus

$$\begin{aligned} W_{n,k} &= \sum_{\mathbf{c} \in B_{n,k}} X_{\mathbf{c}} \\ &= \sum_{\mathbf{c} \in B_{n,k}} \int_{p=0}^1 1_{p \leq X_{\mathbf{c}}} dp \\ &= \int_{p=0}^1 \sum_{\mathbf{c} \in B_{n,k}} 1_{p \leq X_{\mathbf{c}}} dp \end{aligned} \tag{24}$$

$$\begin{aligned}
&= \int_{p=0}^1 |\{\mathbf{c} \in B_{n,k} : p \leq X_{\mathbf{c}}\}| dp \\
&= \int_{p=0}^1 (n^* - \text{rank}(\mathbf{A}_p)) dp.
\end{aligned} \tag{25}$$

Here \mathbf{A}_p is any matrix made up of those columns $\mathbf{c} \in \Omega_{n,k}$ with $X_{\mathbf{c}} \leq p$. And let A_p denote the corresponding hypergraph.

Explanation for (25): Finding a minimum cost basis B can be achieved via a *greedy algorithm*. We first order the columns of $\Omega_{n,k}$ as $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_N, N = \binom{n}{k}$ in increasing order of weight $X_{\mathbf{c}}$. Treating B as a set of columns, we initialise $B = \emptyset$, and for $i = 1, 2, \dots, N$ add \mathbf{c}_i to B if it is linearly independent of the columns of B selected so far. This means that for any $0 \leq p \leq 1$, the number of columns in B with $X_{\mathbf{c}} > p$ must be equal to the co-rank of the set of columns selected before them i.e $B_p = \{\mathbf{c} \in B : X_{\mathbf{c}} \leq p\}$. We claim that B_p is a maximal linear independent subset of the columns of \mathbf{A}_p . If it were not maximal, then another column of \mathbf{A}_p would have been added to B_p by the greedy algorithm.

We obtain $\mathbf{E}W_{n,k}$ in (6) by taking the expectation of (25), using Fubini's theorem to take the expectation inside the integral.

We first argue that

$$\mathbf{E}(W_{n,k}) = \Omega(n^{-(k-2)}). \tag{26}$$

Let $\mathbf{c} = (c_1, \dots, c_n)$, where $c_i \in \{0, 1\}$ denotes the i -th row coordinate of \mathbf{c} . We can bound $W_{n,k}$ from below by $\sum_{i=1}^n \min\{X_{\mathbf{c}} : c_i = 1\} / k$. Let $N = \binom{n}{k}$. The number of ones in a fixed row of $\mathbf{A}_{n,N;k}$ is $L = Nk/n$. The expected minimum of L independent uniform $[0, 1]$ random variables is $1/(L + 1)$. Hence

$$\mathbf{E}(W_{n,k}) \geq \frac{1}{k} \frac{n^2}{k \binom{n}{k} + n}$$

and (26) follows.

We next observe that for c large we have

$$1 - 2ke^{-c} \leq x \leq 1. \tag{27}$$

Indeed, putting $x = 1 - y$ in (1) gives $(1 - y)^{1/(k-1)} = 1 - e^{-c(1-y)}$. We see that if $f(y) = (1 - y)^{1/(k-1)} - (1 - e^{-c(1-y)})$ then $f(0) > 0$ and $f(2ke^{-c}) < 0$ for large c .

Thus for c large we have

$$\frac{c}{k} - \frac{cx^{k/(k-1)}}{k} + (1 - e^{-cx}(1 + cx)) \geq 1 - e^{-cx}(1 + cx) \geq 1 - e^{-99c/100}. \tag{28}$$

Fix some small $\varepsilon > 0$ and let

$$c_\varepsilon = 2 \log 1/\varepsilon. \tag{29}$$

It follows from Theorem 3(i) with $A = k, p = km/(n \binom{n-1}{k-1})$. and (6) that

$$\mathbf{E}(W_{n,k}) \approx \int_{p=0}^{k! \gamma n^{1-k} \log n} (n^* - \mathbf{E}(\text{rank}(\mathbf{A}_p))) dp$$

$$\begin{aligned}
&= \frac{(k-1)!}{n^{k-1}} \int_{c=0}^{k\gamma \log n} (n^* - \mathbf{E}(\text{rank}(\mathbf{A}_{c(k-1)!/n^{k-1}}))) dc \\
&= (I_1 + I_2 + I_3) \frac{(k-1)!}{n^{k-1}},
\end{aligned} \tag{30}$$

where $I_1 = \int_{c=0}^{c_k^*} \dots dc$ and $I_2 = \int_{c_k^*}^{c_\varepsilon} \dots dc$ and $I_3 = \int_{c_\varepsilon}^{k\gamma \log n} \dots dc$.

Since $H_{c/n^{k-1}}$ q.s. has $m \approx cn/k$ edges, it follows from Theorem 1 that

$$I_1 \approx \int_{c=0}^{c_k^*} \left(n^* - \frac{cn}{k} \right) dc \approx c_k^* n \left(1 - \frac{c_k^*}{2k} \right). \tag{31}$$

On the other hand, using the expression for rank from Corollary 2, with $x^{1/(k-1)} = (1 - e^{-cx})$ substituted from (1).

$$I_2 \approx n \int_{c_k^*}^{c_\varepsilon} \left(1 - \left(\frac{c}{k} - \frac{cx^{k/(k-1)}}{k} + (1 - e^{-cx}(1 + cx)) \right) \right) dc \tag{32}$$

$$= n \int_{c_k^*}^{\infty} \left(e^{-cx}(1 + cx(k-1)/k) - \frac{c}{k}(1-x) \right) dc + A_\varepsilon. \tag{33}$$

Using (28) gives

$$|A_\varepsilon| = n \int_{c_\varepsilon}^{\infty} \left(e^{-cx}(1 + cx(k-1)/k) - \frac{c}{k}(1-x) \right) dc \leq n \int_{c_\varepsilon}^{\infty} e^{-99c/100} dc \leq 2\varepsilon n. \tag{34}$$

Theorem 1 as stated holds for $m = O(n)$, and thus cannot be used directly to estimate rank when $m/n \rightarrow \infty$. For I_3 we recall that $\mathbf{C}_2 = \mathbf{C}_2(c)$ denotes the sub-matrix of $\mathbf{A}_{c(k-1)!/n^{k-1}}$ induced by the edges of the 2-core. We then write

$$I_3 \leq \int_{c_\varepsilon}^{k\gamma \log n} (n^* - \mathbf{E}(\text{rank}(\mathbf{C}_2(c)))) dc. \tag{35}$$

We first check the size of $|C_2|$ for $c = c_\varepsilon$. It follows from (1) and (28) that for c large,

$$x^{1/(k-1)} - cx + cx^{k/(k-1)} = 1 - e^{-cx}(1 + cx) \geq 1 - e^{-99c/100}.$$

So, for large enough $c = O(1)$, from (2) we have that w.h.p.

$$|C_2| \geq (1 - o(1))n(1 - e^{-99c/100}).$$

Let $m_\varepsilon = c_\varepsilon n/k$. If we add an edge e with one vertex not in C_2 and the remaining vertices in C_2 then the rank of \mathbf{C}_2 goes up by one. Denote this event by \mathcal{A}_e . Let $\mathbf{C}^* = \mathbf{C}^*(t)$ denote the following submatrix of \mathbf{C}_2 at the time the number of columns is $m_\varepsilon + t$. We let $\mathbf{C}^*(0) = \mathbf{C}_2(c_\varepsilon)$ and we add the column corresponding to e to \mathbf{C}^* only if \mathcal{A}_e occurs. Let X_t denote the rank of $\mathbf{C}^*(t)$, and let $Y_t = n^* - X_t$. Note that X_t is equal to $\text{rank}(\mathbf{C}_2(c_\varepsilon))$ plus the number of columns in $\mathbf{C}^*(t)$ that are not in $\mathbf{C}_2(c_\varepsilon)$, and that $X_t \leq \text{rank}(\mathbf{A}_{m_\varepsilon+t})$.

Note also that $|\text{rank}(\mathbf{A}_{m_\varepsilon+t}) - \text{rank}(\mathbf{A}_{n,p_t,k})| \leq n^{2/3}$ where $p_t = (m_\varepsilon + t)/\binom{n}{k}$. Using (29) we have that $Y_0 \leq (1 + o(1))ne^{-99c_\varepsilon/100} \leq 2\varepsilon n$. Now,

$$\Pr(\mathcal{A}_e) = \frac{Y_t \binom{n-Y_t}{k-1}}{\binom{n}{k}} \geq \frac{kY_t}{2n} \tag{36}$$

and so

$$\mathbf{E}(Y_{t+1} | Y_t) \leq Y_t - \frac{kY_t}{2n}. \tag{37}$$

Let $h = n^{1/2}$ and $u_r = Y_{rh}$. Assume that $n^{9/10} \leq Y_t \leq Y_0$. It follows from (36) and Hoeffding's Theorem [9] that q.s.

$$u_{r+1} \leq u_r - \frac{kh}{3n}u_r = \left(1 - \frac{kh}{3n}\right)u_r$$

and so q.s.

$$u_r \leq \left(1 - \frac{kh}{3n}\right)^r u_0. \tag{38}$$

Going back to (35) we can see that

$$I_3 \leq O(n^{9/10}) + \frac{hu_0}{n} \sum_{r=0}^{\infty} \left(1 - \frac{kh}{3n}\right)^r = O(n^{9/10}) + \frac{3u_0}{k}. \tag{39}$$

Here the final $O(n^{9/10})$ term accounts for only using (37) for $Y_t \geq n^{9/10}$ and for the errors of size $O(n^{2/3})$ introduced in the m model versus the p model of our matrix, see (12), (13).

It follows from (31), (32), (34) and (39) that $I_1 + I_2 + I_3$ are within $O(\varepsilon n)$ of what is claimed in the theorem. Since ε is arbitrary, we can by increasing c_ε , make the value of ε in (29) arbitrarily small and Theorem 5 follows.

3.1 Bounds for finite k

We begin by estimating c_k^* . Let x be as in (1), then going back to the definition (4), we can determine the value of $c_k^* = c(x)$ from

$$c \left(\frac{k-1}{k} \right) x^{\frac{k}{k-1}} - cx + x^{\frac{1}{k-1}} = 0. \tag{40}$$

Solve for c , and put $y = x^{1/(k-1)}$ to give

$$c = \frac{1}{y^{k-2} - ((k-1)/k)y^{k-1}}. \tag{41}$$

Substituting for c via (1) gives

$$y = 1 - \exp \left\{ -\frac{ky}{k - (k-1)y} \right\}. \tag{42}$$

If $x \in (0, 1)$ then $y \in (0, 1)$, and $y \geq x$. We look for solutions of the form $y = 1 - z$. Making this substitution (42) becomes $z = q(z)$ where

$$q(z) = \exp \left\{ -\frac{k(1-z)}{1+(k-1)z} \right\}.$$

Let

$$z = z(\delta) = \frac{\delta}{k - (k-1)\delta}, \tag{43}$$

then (stretching notation somewhat) $q(\delta) = e^{-k(1-\delta)}$. Consider $f(\delta) = z(\delta) - q(\delta)$, then

$$f(\delta) \geq \frac{\delta}{k} \left(1 + \frac{k-1}{k}\delta \right) - e^{-k}e^{k\delta}.$$

Substitute $\delta = \theta ke^{-k}$ to give

$$f(\theta) \geq e^{-k} \left(\theta(1 + \theta(k-1)e^{-k}) - e^{\theta k^2 e^{-k}} \right).$$

The function $k^2 e^{-k}$ in the exponent of the last term is monotone decreasing for $k \geq 2$. Let $\theta = 3/2$, then for $k \geq 4$, it can be checked that $f(\theta, k) > 0$. Now $f(0) < 0$ and so there is a solution to $f(\delta) = 0$ in the interval $(0, \theta ke^{-k})$.

Substitute $y = 1 - z$ into (41) to obtain

$$\frac{c}{k} = \frac{1}{(1-z)^{k-2}(1+(k-1)z)} \tag{44}$$

Lemma 10. (i) Let $\theta = 3/2$, then for $k \geq 4$,

$$k(1 - \theta e^{-k}) \leq c_k^* \leq k. \tag{45}$$

(ii) For $k = 3$, $c_3^* = 2.753813\dots$

(iii) If $k \geq 4$ and $c \geq c_k^*$ then the solution x to (1) satisfies $x \geq 1 - 3ke^{-c}/2$.

Proof. (i) For the upper bound we note that for $k \geq 3$ the denominator of c in (44) is monotone increasing for $z \leq 1/(k-1)^2$ from a value of one when $z = 0$. For the lower bound, as $1/(1-z)^{k-2} > 1 + (k-2)z$, it follows from (44), the definition of z in (43), and $\delta < \theta ke^{-k}$ that

$$\frac{c}{k} > \frac{1 + (k-2)z}{1 + (k-1)z} = 1 - \frac{\delta}{k} > 1 - \theta e^{-k}.$$

(ii) Set $y = \sqrt{x}$ and invert (1) to obtain

$$c = \frac{1}{y^2} \log \frac{1}{1-y}.$$

Inserting this into (41) gives

$$y + \left(\frac{2}{3}y - 1\right) \log \frac{1}{1-y} = 0.$$

This equation was solved numerically to give the following results for y, x, c_3^*

$$y = 0.8834191, \quad x = 0.9399038, \quad c_3^* = 2.753813. \quad (46)$$

(iii) Let $x = 1 - \varepsilon$. We first verify that $\varepsilon \leq 1/c$. Putting $f(\varepsilon) = 1 - \varepsilon - (1 - e^{-c+c\varepsilon})^{k-1}$ we see that $f(0) > 0$ and $f(1/c) < 0$ for $c \geq c_k^*$ as given in (i). If $ay < 1$, then $1 - (1 - y)^a < ay$. As $(k - 1)e^{-c+c\varepsilon} < 1$ for any $\varepsilon < 1 - (\log(k - 1))/c$,

$$f(c^{-1}) = 1 - c^{-1} - (1 - e^{-c+c/c})^{k-1} \leq 1 - c^{-1} - 1 + (k - 1)e^{-c+1}.$$

Now $c(k - 1)e^{-c+1}$ is decreasing as a function of c . And for $k \geq 4$, $k(k - 1)e^{-c+1}$ and $e^{(3k/2)e^{-k}}$ are decreasing as functions in of k . Therefore, for c satisfying (45),

$$c(k - 1)e^{-c+1} < k(k - 1)e^{-(k-1)}e^{(3k/2)e^{-k}} < 1.$$

Let $x = 1 - \varepsilon$, and $\delta = e^{-c+c\varepsilon}$. Rewrite (1) as

$$-\log(1 - \varepsilon) = \varepsilon + \frac{\varepsilon^2}{2} + \dots = (k - 1) \left(\delta + \frac{\delta^2}{2} + \dots \right). \quad (47)$$

It must hold that $\varepsilon \leq (k - 1)\delta$ otherwise the left hand side is greater than the right hand side. Thus, as $\varepsilon < 1/c$,

$$\varepsilon \leq (k - 1)e^{-c+c\varepsilon} \leq (k - 1)e^{-c+1}.$$

A repeated application of this bound, (45) and direct calculation gives

$$\varepsilon \leq (k - 1) \exp \left\{ -c + (k - 1)ce^{-c+1} \right\} \leq (k - 1) \exp \left\{ -c + (k - 1)ke^{1-(1-\theta e^{-k})k} \right\} \leq 3ke^{-c}/2.$$

□

Going back to (31) and using Lemma 10(i), we see that for $k \geq 4$,

$$\frac{kn}{2} \left(1 - \frac{9}{4}e^{-2k} \right) \leq I_1 \leq \frac{kn}{2}. \quad (48)$$

We evaluate I_2 from (32)–(33) in two parts. Firstly, using Lemma 10(iii) for $c \geq c_k^*$,

$$-\frac{3}{2}ce^{-c} \leq -\frac{c}{k}(1 - x) \leq 0. \quad (49)$$

Note also that $1 - 3ke^{-c}/2 \geq 1 - 1/2k$ for $k \geq 4$ and $c \geq c_k^*$. Thus

$$e^{-c} \left(1 + c \frac{(k - 1)(2k - 1)}{2k^2} \right) \leq e^{-cx} \left(1 + cx \frac{k - 1}{k} \right) \leq e^{1/2}e^{-c} \left(1 + \frac{c(k - 1)}{k} \right).$$

For the LHS we replace e^{-cx} by e^{-c} (since $x \leq 1$) and x by $1 - 1/2k$. For the RHS we replace $cx(k-1)$ by $c(k-1)$, and $e^{-cx} = e^{-c+c\varepsilon}$. Using Lemma 10(i) and (iii), as $c^* > 1$, it follows that

$$e^{c\varepsilon} \leq e^{(3k/2)c\varepsilon} \leq e^{(3k/2)c^*e^{-c^*}} \leq e^{1/2}. \quad (50)$$

Adding the contributions from (49) and (50) we find that

$$n \int_{c^*}^{\infty} e^{-c} \left(1 - c \frac{k^2 + 3k - 1}{2k^2} \right) dc \leq I_2 \leq ne^{1/2} \int_{c^*}^{\infty} e^{-c} \left(1 + \frac{c(k-1)}{k} \right) dc.$$

Thus, with the *indefinite integral* $\int e^{-c}(1+Ac) = -e^{-c}(1+A+Ac)$, we get

$$ne^{-c_k^*} \left(\frac{k^2 - 3k + 1}{2k^2} - c_k^* \frac{k^2 + 3k - 1}{2k^2} \right) \leq I_2 \leq ne^{1/2} e^{-c_k^*} \left(\frac{2k-1}{k} + c_k^* \frac{k-1}{k} \right),$$

or more simply

$$-n \frac{k}{2} e^{-c_k^*} \left(1 + \frac{3}{k} \right) \leq I_2 \leq n \frac{k}{2} e^{-c_k^*} 2e^{1/2} \left(1 + \frac{1}{k} \right).$$

Noting that $e^{-c_k^*} \leq 6e^{-k}/5$ for $k \geq 4$, we have

$$n \frac{k}{2} \left(1 - \frac{9}{4} e^{-2k} - \frac{21}{10} e^{-k} \right) \leq I_1 + I_2 \leq n \frac{k}{2} (1 + 3e^{1/2} e^{-k}).$$

Thus, for some ε_k , $|\varepsilon_k| \leq 5$,

$$I_1 + I_2 = n \frac{k}{2} (1 + \varepsilon_k e^{-k}).$$

4 Open questions

Q1 The formula for the cost of a minimum weight basis when $k \geq 3$ given by Theorem 5 is asymptotically accurate, but lacks the elegance of the case where $k = 2$. Can the expression be simplified for say, $k = 3$?

Q2 The $\zeta(3)$ result of [8] was generalised quite substantially to consider minimum weight spanning trees of d -regular graphs, when d is large, see [4]. In the context of $\mathbf{A}_{n,m;k}$, this suggests that we consider the case where each row has exactly d ones. Here we can study the rank as well as $W_{n,k}$.

References

- [1] J. Altschuler and E. Yang, Inclusion of Forbidden Minors in Random Representable Matroids, *Discrete Mathematics* 340 (2017) 1553-1563.

- [2] P. Ayre, A. Coja-Oghlan, P. Gao and N Müller, The satisfiability threshold for random linear equations, [arXiv:1710.07497](https://arxiv.org/abs/1710.07497).
- [3] N. Bansal, R.A. Pendavingh and J.G. van der Pol, On the number of matroids, *Combinatorica* 35 (215) 253-277.
- [4] A. Beveridge, A. M. Frieze and C. J. H. McDiarmid, Minimum length spanning trees in regular graphs, *Combinatorica* 18 (1998) 311-333.
- [5] C. Cooper, On the rank of random matrices, *Random Structures and Algorithms* 16 (2000) 209-232.
- [6] C. Cooper, The cores of random hypergraphs with a given degree sequence, *Random Structures and Algorithms* 25 (2004) 353-375.
- [7] C. Cooper, A.M. Frieze and W. Pegden, Minors of a random binary matroid, [arXiv:1612.02084v5](https://arxiv.org/abs/1612.02084v5).
- [8] A.M. Frieze, On the value of a random minimum spanning tree problem, *Discrete Applied Mathematics* 10 (1985) 47-56.
- [9] W. Hoeffding, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58 (1963) 13-30.
- [10] S. Janson, The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph, *Random Structures and Algorithms* 7 (1995) 337-355.
- [11] W. Kordecki and A. Lyczkowska-Hanćkowiak, Exact Expectation and Variance of Minimal Basis of Random Matroids, *Discussiones Mathematicae Graph Theory* 33 (2013) 277-288.
- [12] M. Molloy, Cores in random hypergraphs and random formulas, *Random Structures and Algorithms* 27 (2005) 124-135.
- [13] J. Oxley and D. Kelly, On random representable matroids, *Studies in Applied Mathematics* 71 (1984) 181-205.
- [14] J. Oxley, L. Lowrance, C. Semple and D. Welsh, On properties of almost all matroids, *Advances in Applied Mathematics* 50 (2013) 115-124.
- [15] B. Pittel and G. Sorkin, The Satisfiability Threshold for k -XORSAT, *Combinatorics, Probability and Computing* 25 (2016) 238-268.