Rank of the vertex-edge incidence matrix of r-out hypergraphs

Colin Cooper
Department of Informatics
King's College
London WC2B 4BG
England

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

January 12, 2022

Abstract

We consider the rank of a class of sparse Boolean matrices of size $n \times n$. In particular, we show that the probability that such a matrix has full rank, and is thus invertible, is a positive constant with value about 0.2574 for large n.

The matrices arise as the vertex-edge incidence matrix of 1-out 3-uniform hypergraphs. The result that the null space is bounded in expectation, can be contrasted with results for the usual models of sparse Boolean matrices, based on the vertex-edge incidence matrix of random k-uniform hypergraphs. For this latter model, the expected co-rank is linear in the number of vertices n, [5], [8].

For fields of higher order, the co-rank is typically Poisson distributed.

1 Introduction

For positive integers $r \geq 1,\ s \geq 2$, let $\boldsymbol{M}(s, r, n)$ be the space of $n \times rn$ matrices with entries generated in the following manner. For each $i = 1, ..., n$ there are r columns $C_{i,j},\ j = 1, ..., r$. Each column $C_{i,j}$ has a unit entry in row i, and $s-1$ other unit entries, in rows chosen randomly with replacement from $[n]$, or without replacement from $[n] - \{i\}$, all other entries in the column being zero. In general we consider the arithmetic on entries in the matrix,

*Research supported in part by NSF Grant DMS1661063

(and thus the evaluation of linear dependencies), to be over $GF(2)$. If so, in the "with replacement case", if two unit entries coincide the entry is set to zero. When $r = 1$, the matrix consists of an identity matrix plus $s-1$ random units in each column.

If $s = 2$, and entries are chosen without replacement, M is the vertex-edge incidence matrix of the random graph $G_{r-\text{out}}(n)$. This model of random graphs has been extensively studied, and is known to be r-connected for $r \geq 2$, Fenner and Frieze [10], to have a perfect matching for $r \geq 2$, Frieze [11], and to be Hamiltonian for $r \geq 3$, Bohman and Frieze [4]. If $s \geq 3$ we are considering r-out, s-uniform hypergraphs. Random Boolean matrices based on the vertex-edge incidence matrix of s-uniform hypergraphs where the columns (edges) are chosen i.i.d. from all columns with s ones were studied by Cooper, Frieze and Pegden, [8]. A very general paper by Coja-Oghlan, Ergür, Gao, Hetterich and Rolvien, [5], gives the limiting rank in this latter model for a wide range of assumptions on the distribution of non-zero entries in the rows and columns. The fundamental difference between the r-out model of random matrices, and those of [5], [8] is the presence of an $n \times n$ identity matrix as a sub-matrix (in the without replacement case).

We will use ρ to denote the (row) rank of our matrices and then the co-rank is $n - \rho$. If the field is $GF(2)$, $\boldsymbol{x} \in \{0,1\}^n$ is a linear dependency (*dependency* for short) if $\boldsymbol{x}M = 0$. Let $|\boldsymbol{x}| = |\{j : x_j = 1\}|$. We say that a set of rows $D \subseteq [n]$ is a dependency if $D = \{j : x_j = 1\}$ for some dependency $\boldsymbol{x}$. An ℓ-dependency is one where $|\boldsymbol{x}| = \ell$ or $|D| = \ell$.

Of particular interest is the case $r = 1$ which gives $n \times n$ Boolean matrices. We will show that over $GF(2)$, for $r = 1, s = 3$, the linear dependencies among the rows of M are w.h.p. either small bounded in expectation or large (of size about $n/2$), and the distributions of these dependencies are somewhat entangled. Estimating the interaction between small and large dependencies in matrices from $\boldsymbol{M}(3, 1, n)$ is the main problem we solve.

For $r = 1, s = 3$, define a Poisson parameter ϕ for small dependencies. The value of ϕ differs between the "with replacement" ϕ_R, and "without replacement" models $\phi_{\overline{R}}$ as follows:

$$\phi_R = \sum_{\ell \geq 1} \frac{1}{\ell}(2e^{-2})^\ell \sum_{j=0}^{\ell-1} \frac{\ell^j}{j!}, \qquad \phi_{\overline{R}} = \sum_{\ell \geq 2} \frac{1}{\ell}(2e^{-2})^\ell \sum_{j=0}^{\ell-2} \frac{\ell^j}{j!}. \tag{1}$$

The numeric values are $\phi_R \approx 0.5215$, and $\phi_{\overline{R}} \approx 0.1151$, where $a \approx b$ means approximately equal.

Let

$$P(\sigma, \lambda) = \left(\frac{1}{2}\right)^{\lambda(\lambda+\sigma)} \frac{1}{\prod_{j=1}^{\lambda}\left(1 - \left(\frac{1}{2}\right)^j\right)} \prod_{j=1}^{\infty}\left(1 - \left(\frac{1}{2}\right)^{\lambda+\sigma+j}\right). \tag{2}$$

The quantity $P(\sigma, \lambda)$ is the limiting value of $\mathbb{P}(\lambda \mid \sigma)$ of the conditional probability of

$\lambda = d - \sigma$ given σ, where σ is the dimension of the space induced by *small* dependencies and d the dimension of the space induced by all dependencies.

Theorem 1. *Let the matrix M be chosen u.a.r. from $\boldsymbol{M}(3,1,n)$. Let $d \geq 0$ be integer. Over $GF(2)$, the limiting probability M has co-rank d is given by*

$$\lim_{n\to\infty} \mathbb{P}(\text{co-rank}(M) = d) = e^{-\phi} \sum_{\sigma=0}^{d} \frac{\phi^{\sigma}}{\sigma!} \, P(\sigma, d-\sigma). \tag{3}$$

In particular,

$$\mathbb{P}(\text{rank}(M) = n) \sim e^{-\phi} P(0,0) = e^{-\phi} \prod_{j=1}^{\infty} \left(1 - \left(\frac{1}{2}\right)^{j}\right).$$

Theorem 1 differs from many previous results on sparse random Boolean matrices. The co-rank (dimension of the null space) is bounded in expectation, and the matrix is invertible with probability $e^{-\phi}P(0,0) \approx 0.2574$ in the without replacement model. The bounded co-rank given by Theorem 1 can be contrasted with results for the edge-vertex incidence matrix of random hypergraphs, ([5], [8]), where the expected co-rank is linear in the number of vertices n, and the probability of a full rank matrix is exponentially small.

The matrices $\boldsymbol{M}(3,1,n)$ exhibit a gap in the size of the dependencies (small or large), which we next explain.

Theorem 2. *Let M be chosen u.a.r. from $\boldsymbol{M}(3,1,n)$, then w.h.p. either (i) a dependency $\boldsymbol{x}$ is* small *i.e. $|\boldsymbol{x}| \leq \omega$ where $\omega \to \infty$ slowly or (ii) $\boldsymbol{x}$ is* large *i.e. $|\boldsymbol{x}| = n/2 + O(\sqrt{n \log n})$.*

A gap property in solutions to random XOR-SAT systems over $GF(2)$ was previously observed by Achiloptas and Molloy [1], and by Ibrahimi, Kanoria, Kraning and Montanari [13]. They found that the Hamming distance between XOR-SAT solutions was either $O(\log n)$ or at least αn; where n is the number of variables. In our case, large dependencies have intersection about $n/4$ (see Section 4), giving a precise value of α.

A dependency $\boldsymbol{x}$ is *fundamental* if there is no other dependency $\boldsymbol{y} \neq \boldsymbol{x}$ such that $\boldsymbol{y} \leq \boldsymbol{x}$, componentwise. We will prove in Section 2 that the number Z of fundamental small dependencies is asymptotically distributed as $Po(\phi)$ i.e. Poisson with mean ϕ. The quantity $P(\sigma, \lambda)$ in (3) is the limiting probability that small dependencies span a space of dimension σ, and large dependencies increase the co-rank by λ.

The value of $P(0,k)$ given in (2), is the same as the value $\pi(k)$ given in (39). The probability distribution defined by π was previously observed in a model of random matrices over $GF(2)$ in which the entries $m_{i,j}$ are i.i.d Bernoulli random variables with $\mathbb{P}(m_{i,j} = 1) = p$. For a wide range of p the distribution of dimension k of the null space is given by $\pi(k)$. The result was proved by Kovalenko et al., [14] for $p = 1/2$, and extended to the range $\min(p(n), 1-p(n)) \geq (\log n + c(n))/n$, (where $c(n) \to \infty$ slowly) by Cooper [6]. A similar distributional result holds for the model of random matrices over the finite field $GF(q)$, see Cooper [7]. Here the non-zero entries $\alpha \in GF(q) \setminus \{0\}$ are independently and uniformly distributed with $\mathbb{P}(m_{i,j} = \alpha) = p/(q-1)$. The distribution of co-rank $\pi_q(k)$ equivalent to $\pi(k) = \pi_2(k)$ in (39) is obtained by replacing the $(1/2)$ terms in (39) by $(1/q)$.

Finally we mention some related cases for r-out s-uniform hypergraphs. For $r = 1$ and $s = 2$, M has expected rank $\sim n - (\log n)/2$. This is because the expected number of components in a random mapping is $\sim (1/2) \log n$, (see e.g., [12]). Note: For s even, the rows of M add to zero modulo 2. The following theorem is immediate from the proof of Theorem 1.

Theorem 3. *If $r \geq 2$ and $s = 2,\ 3$, then M has rank $n^* = n - \mathbb{1}_{\{s=2\}}$, w.h.p.*

The proof of Theorem 3, and results for finite fields of character $q \geq 3$ can be found in [9].

Notation: Apart from $O(\cdot), o(\cdot), \Omega(\cdot)$ as a function of $n \to \infty$, we use the notation $A_n \sim B_n$ if $\lim_{n\to\infty} A_n/B_n = 1$. The symbol $a \approx b$ indicates approximate numerical equality due to decimal truncation. The notation $\omega(n)$ describes a function tending to infinity as $n \to \infty$. The expression *with high probability* (w.h.p.), means with probability $1 - o(1)$, where the $o(1)$ is a function of n, which tends to zero as $n \to \infty$.

Outline of the proof for $GF(2)$ with $r = 1, s = 3$

Because the proofs are rather technical, *we give a detailed proof in the "with replacement" model*. For brevity, we omit the proof that the results are also valid in the "without replacement" model in this paper; the proof can be found in [9].

We refer to the rows of M as $M_i, i \in [n]$ and to the columns as $C_j, j \in [n]$. By a set of rows S, we mean the set of rows $M_i, i \in S$. A set of rows with indices L is linearly dependent (zero-sum) if $\sum_{i\in L} M_i = 0 (\text{mod } 2)$. A linear dependence L is *small* if $|L| \leq \omega$, where $\omega = \omega(n)$ is a function tending slowly to infinity with n. A linear dependence L is *large* if $|L| = (n/2)(1 + O(\sqrt{\log n/n}))$. As part of our proof, we show that w.h.p. there are no other sizes of dependency. A set of zero-sum rows L is *fundamental*, if L contains no smaller zero-sum set and L is disjoint from all other zero-sum sets. The zero-sum sets of size about $n/2$ are not disjoint. We count k-sequences of large dependencies with a property we call *simple*. Many of the problems with the proofs arise because large dependencies are not disjoint, and are conditioned by the simultaneous presence of small dependencies in M.

We next outline the main steps in the proof of Theorem 1.

1. In Section 2 we prove that the number Z of small fundamental dependencies has factorial moments $\mathbf{E}(Z)_k \sim \phi^k$, where ϕ is given by (1). Thus Z is asymptotically

Poisson distributed and

$$\mathbb{P}\left(M \text{ has } i \text{ small fundamental linear dependencies}\right) \sim \frac{\phi^i}{i!}e^{-\phi}.$$

2. For $M \in \boldsymbol{M}(3,1,n)$ w.h.p. any fundamental sets of zero-sum rows of M are either small (of size $\ell \leq \omega$) or large (of size $\ell = (n/2)(1+O(\sqrt{\log n/n}))$). This is proved in Section 3.

3. In Section 5 we discuss *simple* sequences of large dependencies, and in Section 6 we estimate the moments of these sequences and determine their interaction with small dependencies.

4. In Section 7 we estimate the number of simple sequences, conditional on the the number of small fundamental dependencies. This leads to an approximate set of linear equations whose solution completes the proof of Theorem 1.

2 Small dependencies in $GF(2)$: with replacement

Notation For $1 \leq k \leq \omega$, where $\omega \to \infty$ arbitrarily slowly with n, let $X_k(M)$ or $Y_k(M)$ denote the number of index sets of k-dependencies in M. A k-dependency is *small* if $k \leq \omega$ and we use Y_k when $k \leq \omega$ and use X_k when $k \sim n/2$. We will show that for other values of k, $X_k = 0$ w.h.p. We also use $Z_d, d \leq \omega$ to denote the number d of fundamental (minimal) dependent sets among the rows of M.

We first consider dependencies with $s = o(n^{1/2})$ rows. For $S \subseteq [n]$, let $\mathcal{F}(S)$ denote the event that the rows corresponding to S are dependent. Let Y_s denote the number of s-set dependencies.

Lemma 4. *If $|S| = s = o(n^{1/2})$ then*

$$\mathbb{P}(\mathcal{F}(S)) \sim \left(\frac{2s}{n}\right)^s e^{-s}. \tag{4}$$

If $\omega \to \infty$, $\omega \leq s = o(n^{1/2})$ then $Y_s = 0$ w.h.p.

Proof. Suppose that $s = o(n^{1/2})$ and $S = [s]$. Then,

$$\begin{aligned}\mathbb{P}(\mathcal{F}(S)) &= \left(2\left(\frac{s}{n}\right)\left(\frac{n-s}{n}\right)\right)^s \left(\left(\frac{s}{n}\right)^2 + \left(\frac{n-s}{n}\right)^2\right)^{n-s} \\ &\sim \left(\frac{2s}{n}\right)^s e^{-2s}, \qquad \text{using } s = o(\sqrt{n}). \end{aligned} \tag{5}$$

Explanation: The probability that exactly one of the two random choices in a column of S lies in a row of S is $2\left(\frac{s}{n}\right)\left(\frac{n-s}{n}\right)$. The probability that both or neither of the two random choices in a column of $[n]\setminus S$ lies in a row of S is $\left(\frac{s}{n}\right)^2+\left(\frac{n-s}{n}\right)^2$.

This verifies (4). It follows that

$$\mathbf{E}\left(Y_s\right)\sim\binom{n}{s}\left(\frac{2s}{n}\right)^s e^{-2s}\sim\frac{(2s)^s e^{-2s}}{s!},$$

As $\mathbf{E}\,Y_{s+1}/\mathbf{E}\left(Y_s\right)\sim 2/e$ we have that $\mathbf{E}\,Y_\omega=e^{-\Omega(\omega)}$ and so w.h.p. there are no dependencies with $\omega\le s=o(n^{1/2})$. □

Define $\sigma_s,\ \kappa_s$ by

$$\sigma_s=\sum_{j=0}^{s-1}\frac{s^j}{j!},\qquad\text{and}\qquad\kappa_s=\frac{(s-1)!}{s^s}\sigma_s. \tag{6}$$

For $S\subseteq[n]$, let $\mathcal{F}^*(S)$ denote the event that the rows corresponding to S form a fundamental dependency. The next three lemmas deal with small fundamental dependencies.

Lemma 5. $\mathbb{P}(\mathcal{F}^*(S)\mid\mathcal{F}(S))=\kappa_s$.

Proof. The rows of the dependency S consist of an $s\times s$ sub-matrix $M_{S,S}$ and a zero $(s\times n-s)$ sub-matrix. For $i\in S$, if $M_{i,i}=1$, then w.h.p. there is a unique entry $M_{j,i}=1$ which gives rise to an *edge* (i,j). If $M_{i,i}=0$ we regard this as a loop (i,i). Thus $M_{S,S}$ is the incidence matrix of a random functional digraph D_S, and S is fundamental iff the underlying graph of D_S is connected. For $s\ge 1$, $\mathbb{P}(D_S$ is connected$)=\kappa_s$ (see e.g., [2] or [12]). □

Lemma 6. *Small fundamental dependent sets of M are pairwise disjoint, w.h.p.*

Proof. Let S,T be two small fundamental zero-sum row sets with a non-trivial intersection $C=S\cap T$ and differences $A=S\setminus T,\ B=T\setminus S$, where $A\cup B\neq\emptyset$. Suppose $A\neq\emptyset$. As the functional digraphs D_S,D_T are connected, one of the following events must occur. Either (i) some column of C has two non-zero entries in the rows of $S\cup T$; or (ii) some column j of A has a non-zero entry in the rows of C. The latter is not possible as then a column of S has a non-zero entry in the rows of T. Let $k=|S\cup T|$. The former has probability at most

$$\sum_{k=2}^{2\omega}\binom{n}{k}k\left(\frac{k}{n}\right)^{k-1}\left(\frac{k}{n}\right)^2=o(1). \tag{7}$$

□

Given this lemma we can now prove a Poisson distribution.

Lemma 7. *The number Z of small fundamental dependent sets among the rows of M is asymptotically Poisson distributed with parameter ϕ_R, and thus*

$$\mathbb{P}(Z = d) \sim \frac{\phi_R^d}{d!} e^{-\phi_R}. \tag{8}$$

Proof. Fix $S \subseteq [n]$ and let $S_1, \dots, S_d$ be a partition of S with $|S_i| = s_i, i = 1, 2, \dots, d$. Let $P(s_1, \dots, s_d)$ be the probability that each $S_i, i = 1, 2, \dots, d$ is a fundamental set, given that S is a dependency. Thus,

$$P(s_1, \dots, s_d) = \frac{(s_1)^{s_1} \cdots (s_d)^{s_d}}{s^s} \prod_{i=1,\dots,d} \mathbb{P}(D_{S_i} \text{ connected}) = \frac{1}{s^s} \prod_{i=1}^{d} (s_i - 1)! \sigma_{s_i}.$$

Explanation: the factor $\frac{(s_1)^{s_1} \cdots (s_d)^{s_d}}{s^s}$ is the conditional probability that the random choices for columns with index in S_i are in rows with index in S_i.

Thus, using (4), we see that

$$\begin{aligned}
\mathbf{E}\,(Z)_d \sim & \sum_{s \geq 1} \frac{(2s)^s}{s!} e^{-2s} \sum_{s_1 + \cdots + s_d = s} \binom{s}{s_1, \dots, s_d} P(s_1, \dots, s_d) && (9) \\
= & \sum_{s \geq 1} \sum_{s_1 + \dots + s_d = s} \prod_{i=1}^{d} (2e^{-2})^{s_i} \frac{1}{s_i} \sigma_{s_i} \\
= & \left(\sum_{s \geq 1} \frac{1}{s} (2e^{-2})^s \sigma_s \right)^d \\
= & \phi_R^d. && (10)
\end{aligned}$$

Thus, by the method of moments, the number of small disjoint fundamental zero-sum sets Z tends tend to a Poisson distribution with parameter ϕ_R. □

3 Large zero-sum sets: First moment calculations

Define an index set J_a as follows,

$$J_a = \{n/2 - \sqrt{an \log n} \leq \ell \leq n/2 + \sqrt{an \log n}\} \text{ and } \overline{J}_a = [n] \setminus J_a,\ a \geq 0. \tag{11}$$

Lemma 8. *(***Large linearly dependent sets.)** *Let X_ℓ denote the number of ℓ-dependencies among the rows of M.*

(i) $\sum_{\ell\in J_1}\mathbf{E}\,X_\ell \sim 1$.

(ii) *Let* $F=[n]\backslash([\omega]\cup J_1)$, *where* $\omega\to\infty$ *arbitrarily slowly with* n. *Then* $\sum_{\ell\in F}\mathbf{E}\,X_\ell = o(1)$.

Proof. From (5), the expected number of dependencies of size ℓ is

$$\mathbf{E}\,X_\ell = \binom{n}{\ell}\left(2\left(\frac{\ell}{n}\right)\left(\frac{n-\ell}{n}\right)\right)^{\ell}\left(\left(\frac{\ell}{n}\right)^2+\left(\frac{n-\ell}{n}\right)^2\right)^{n-\ell}.$$

We next approximate the expression for $\mathbf{E}\,X_\ell$. We note the following expansion.

$$(1+x)\log(1-x^2)+(1-x)\log(1+x^2) = -2\left(x^3+\frac{x^4}{2}+\frac{x^7}{3}+\sum_{k\ge 4}\mathbb{1}_{\{k\text{ even}\}}\frac{x^{2k}}{k}\left(1+\frac{kx^3}{k+1}\right)\right). \tag{12}$$

We write $\mathbf{E}\,X_\ell=\binom{n}{\ell}\Phi_\ell^n$, $\ell=(n/2)(1+\varepsilon)$, where

$$\begin{aligned}
\Phi_\ell &= \left(\frac{1-\varepsilon^2}{2}\right)^{\frac{(1+\varepsilon)}{2}}\left(\left(\frac{1+\varepsilon}{2}\right)^2+\left(\frac{1-\varepsilon}{2}\right)^2\right)^{\frac{(1-\varepsilon)}{2}}\\
&=\frac{1}{2}(1-\varepsilon^2)^{\frac{(1+\varepsilon)}{2}}(1+\varepsilon^2)^{\frac{(1-\varepsilon)}{2}}\\
&=\frac{1}{2}\exp\left\{\frac{1}{2}\left((1+\varepsilon)\log(1-\varepsilon^2)+(1-\varepsilon)\log(1+\varepsilon^2)\right)\right\}\\
&=\frac{1}{2}\exp\left\{-\left(\varepsilon^3+\frac{\varepsilon^4}{2}+\frac{\varepsilon^7}{3}+\sum_{k\ge 4}\mathbb{1}_{\{k\text{ even}\}}\varepsilon^{2k}\left(\frac{1}{k}+\frac{\varepsilon^3}{k+1}\right)\right)\right\}\\
&=\frac{1}{2}\exp\left\{-\left(\varepsilon^3+\frac{\varepsilon^4}{2}+\varepsilon_7\right)\right\},
\end{aligned}\tag{13}$$

where $|\varepsilon_7|\le 2|\varepsilon|^7/3$ for sufficiently small ε.

Also for $\ell=(n/2)(1+\varepsilon)$, $|\varepsilon|<1$,

$$\binom{n}{\ell} = \left(1+O\left(\frac{1}{n}\right)\right)\frac{2^n}{\sqrt{2\pi n(1-\varepsilon^2)}}\exp\left(-n\left(\frac{\varepsilon^2}{2}+\frac{\varepsilon^4}{12}+\varepsilon_6\right)\right), \tag{14}$$

where $|\varepsilon_6|\le|\varepsilon|^6/10$.

Case 1: $\ell\in J_1$ **.** From (14) with $|\varepsilon|=2\sqrt{(\log n)/n}$ we have

$$\frac{1}{2^n}\sum_{\ell\notin J_1}\binom{n}{\ell} = O(1/n^{5/2}),$$

so that

$$\frac{1}{2^n}\sum_{\ell\in J_1}\binom{n}{\ell} = 1 - O(1/n^{5/2}).$$

Using (13), for $\ell \in J_1$, ${\Phi_\ell}^n = e^{\Theta(n\varepsilon^3)}/2^n$. Then, as $n\varepsilon^3 = O(\log^{3/2} n/\sqrt{n})$,

$$\sum_{\ell\in J_1} \mathbf{E}\, X_\ell = \sum_{\ell\in J_1}\binom{n}{\ell}\frac{1}{2^n}e^{\Theta(n\varepsilon^3)} = 1 + o(1).$$

For future reference, we note that for $|\varepsilon| < c < 1$,

$$\begin{aligned}
\mathbf{E}\, X_\ell =& \binom{n}{\ell}\frac{1}{2^n}\exp\left\{-n\left(\varepsilon^3 + \frac{\varepsilon^4}{2} + \varepsilon_7\right)\right\} \\
=& \frac{(1+o(1))}{\sqrt{2\pi n(1-\varepsilon^2)}}\exp\left\{-n\left(\frac{\varepsilon^2}{2} + \varepsilon^3 + \frac{\varepsilon^4}{2} + \frac{\varepsilon^4}{12} + \varepsilon_6 + \varepsilon_7\right)\right\} \\
=& \frac{(1+o(1))}{\sqrt{2\pi n(1-\varepsilon^2)}}\exp\left\{-\frac{n\varepsilon^2}{2}\left((1+\varepsilon)^2 + \frac{\varepsilon^2}{6} + O(\varepsilon^4)\right)\right\}. \qquad (15)
\end{aligned}$$

Case 2: $\ell \in F$**.** Write $F = [n] \setminus ([\omega] \cup J_1)$ as $F = F_1 \cup F_2 \cup F_3$ where $F_1 = \{\omega, \ldots, 3n/10\}$, $F_2 = \{7n/10, \ldots, n\}$ and $F_3 = F \setminus (F_1 \cup F_2)$. Thus, for $\ell \in F_3$, $\ell = (n/2)(1+\varepsilon)$ where $-2/5 \le \varepsilon \le -\sqrt{(2\log n)/n}$ or $\sqrt{(2\log n)/n} \le \varepsilon \le 2/5$.

Case $\ell \in F_1$. For sufficiently large n, Stirling's approximation implies that

$$\binom{n}{\ell} \le \frac{n^n}{\ell^\ell(n-\ell)^{n-\ell}},$$

so for some constant C (in both with and without replacement models)

$$\mathbf{E}\, X_\ell \le \frac{Cn^n}{\ell^\ell(n-\ell)^{n-\ell}}\left(2\left(\frac{\ell}{n}\right)\left(\frac{n-\ell}{n}\right)\right)^\ell\left(\left(\frac{\ell}{n}\right)^2 + \left(\frac{n-\ell}{n}\right)^2\right)^{n-\ell}. \qquad (16)$$

Continuing with this expression, using $\ell = \lambda n$ for $\lambda < 1/2$,

$$\begin{aligned}
\mathbf{E}\, X_\ell \le& C\left(\frac{2^\lambda}{\lambda^\lambda(1-\lambda)^{1-\lambda}}\lambda^\lambda(1-\lambda)^\lambda(\lambda^2 + (1-\lambda)^2)^{1-\lambda}\right)^n \\
=& C\left(2^\lambda(1-\lambda)^\lambda\left(1-\lambda+\frac{\lambda^2}{1-\lambda}\right)^{1-\lambda}\right)^n \\
\le& C\left(2^\lambda(1-\lambda)^\lambda e^{-\lambda(1-\lambda)+\lambda^2}\right)^n \\
=& C\left(2(1-\lambda)e^{-1+2\lambda}\right)^{\lambda n} \\
=& C[g(\lambda)]^{\lambda n}.
\end{aligned}$$

The function $g(\lambda)$ is strictly concave and has a unique maximum at $\lambda = 1/2$ with $g(1/2) = 1$. For $\lambda \leq 3/10$, $g(\lambda) \leq g(3/10) = (7/5)e^{-2/5} < 1$ so that

$$\sum_{\ell \in F_1} \mathbf{E}\, X_\ell \leq C \sum_{\ell \in F_1} g(3/10)^\ell = o(1).$$

Case $\ell \in F_2$. Referring to (15), the function $h(\varepsilon) = (\varepsilon^2/2)((1+\varepsilon)^2 + \varepsilon^2/6 + \varepsilon_6 + \varepsilon_7)$ satisfies $h(\varepsilon) > 2/25$ for $\varepsilon \geq 2/5$, and so

$$\sum_{\ell \in F_2} \mathbf{E}\, X_\ell \leq \sum_{\ell \in F_2} e^{-\Omega(n)} = o(1).$$

Case $\ell \in F_3$. For $\sqrt{(2\log n)/n} \leq |\varepsilon| \leq \sqrt{(25\log n)/n}$, the function $h(\varepsilon) \geq (1-o(1))(\log n)/n$. Let F_{3a} be the values of ℓ in this range

$$\sum_{\ell \in F_{3a}} \mathbf{E}\, X_\ell = O(\sqrt{n\log n})/n^{1-o(1)}) = o(1/n^{1/3}).$$

Let $F_{3b} = F_3 \setminus F_{3a}$. Then $\varepsilon^2/2 \geq (25/2)(\log n)/n$, and $(1+\varepsilon)^2 + \varepsilon^2/6 + \varepsilon_6 + \varepsilon_7 > 9/25$. Referring to (15),

$$\sum_{\ell \in F_{3b}} \mathbf{E}\, X_\ell = O(n)/n^4 = o(1/n^3).$$

□

4 Higher moments of large zero-sum sets: Background

Let $A \oplus B$ denote the symmetric set difference of the sets A and B. Thus $A \oplus B = (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$. Suppose that, over $GF(2)$, the rows $M[i], i \in A$ indexed by A are zero-sum, thus $\boldsymbol{z}_A = \sum_{i \in A} M[i] = \mathbf{0}$. Let B be another set such that $\boldsymbol{z}_B = \mathbf{0}$. We can write $\boldsymbol{z}_A = \boldsymbol{z}_{A\setminus B} + \boldsymbol{z}_{A \cap B}$ and $\boldsymbol{z}_B = \boldsymbol{z}_{B \setminus A} + \boldsymbol{z}_{A \cap B}$. Adding these two sets of rows modulo 2 has the effect of canceling the intersection $A \cap B$. Thus (i) $\boldsymbol{z}_A + \boldsymbol{z}_B = 0$, whether $\boldsymbol{z}_{A\cap B}$ is itself zero-sum or not; and (ii) $\boldsymbol{z}_A + \boldsymbol{z}_B = \boldsymbol{z}_{A \oplus B}$.

Recall that a set of zero-sum rows is fundamental if it contains no smaller zero-sum set of rows. For small sets we were able to count fundamental dependencies directly. We have to adopt an alternative strategy for large zero-sum sets. We use an approach similar to the one given in [6]. We count *simple* sequences of large linearly dependent row sets $B = (B_1, ..., B_k)$, $k \geq 1$ constant, and where $|B_i| \in J_1$ so that $|B_i| \sim n/2$. A k-tuple of large dependent sets $B = (B_1, ..., B_k)$ is simple, if for all sequences $(j_1 < j_2 < ... < j_l)$ and $(1 \leq l \leq k)$ the set differences satisfy

$$|B_{j_1} \oplus B_{j_2} \oplus \cdots \oplus B_{j_l}| \in J_1 \tag{17}$$

For any given matrix M there is a largest k such that $B_1, ..., B_k$ are simple. In which case, we say k is *maximal* and $B_1, ..., B_k$ is a *maximal simple sequence*.

Let $V(M) = \{\emptyset\} \cup \{B : B \text{ is zero-sum in } M\}$, then $(V(M), \oplus)$ is a vector space over GF_2 under the convention that $0 \cdot B = \emptyset,\ 1 \cdot B = B$. In $V(M)$ a simple sequence $(B_1, ..., B_k)$ is an ordered basis for a subspace S of dimension k.

Given k linearly dependent sets of rows with index sets $B_1, \cdots, B_k$, there are 2^k intersections of these sets and their complements. For each $\boldsymbol{x} = (x_1, \cdots, x_k)$, $\boldsymbol{x} \in \{0,1\}^k$ we let $I_{\boldsymbol{x}} = \cap_{i=1,...,k} B_i^{(x_i)}$ where $B_i^{(0)} = \overline{B}_i = [n] \setminus B_i$ and $B_i^{(1)} = B_i$. The index sets $I_{\boldsymbol{x}}$ are disjoint by definition and their union (including $\boldsymbol{x}_0 = (0, \cdots, 0)$) is $[n]$.

Next for $\boldsymbol{x} \in \{0,1\}^k$ let $B(\boldsymbol{x}) = \bigoplus_{i:x_i=1} B_i$. Let $K = 2^k - 1$. Let U be a $K \times K$ matrix indexed by $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^k$, $\boldsymbol{x}, \boldsymbol{y} \neq 0$; with entries $U(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $I_{\boldsymbol{y}} \subseteq B(\boldsymbol{x})$, and $U(\boldsymbol{x}, \boldsymbol{y}) = 0$ otherwise. In summary,

$$\text{Row index } \boldsymbol{x} = (x_1, x_2, \ldots, x_k) \text{ is the indicator vector for } B(\boldsymbol{x}) = \bigoplus_{i:x_i=1} B_i,$$
$$\text{Column index } \boldsymbol{y} = (y_1, y_2, \ldots, y_k) \text{ is the indicator vector for } I_{\boldsymbol{y}} = \bigcap_{i=1,\ldots,k} B_i^{(y_i)}.$$

The row of U representing the set $B(\boldsymbol{x})$ is formed by adding the rows of those sets B_i such that $x_i = 1$ in $\boldsymbol{x}$; the addition being over $GF(2)$. Thus $B(\boldsymbol{x})$ is the union of the sets $I_{\boldsymbol{y}}$, where $y_i = 1$ for an odd number of those sets B_i where $x_i = 1$. This can be seen inductively by generating B_1, $B_1 \oplus B_2$, $(B_1 \oplus B_2) \oplus B_3$ etc. in the given order. In summary $U(\boldsymbol{x}, \boldsymbol{y}) = 1$ iff both $x_i = 1$ and $y_i = 1$ for an odd number of indices i, and thus, over $GF(2)$,

$$U(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{k} x_i y_i. \tag{18}$$

Our aim is to use U, treated as a real matrix to show that w.h.p. $|I_{\boldsymbol{x}}| \sim n/2^k$ for every $\boldsymbol{x}$. We do this by observing that given the characterisation $U(\boldsymbol{x}, \boldsymbol{y}) = 1_{I_{\boldsymbol{y}} \subseteq B(\boldsymbol{x})}$, the vector $(|I_{\boldsymbol{x}}|, \boldsymbol{x} \in \{0,1\}^k, \boldsymbol{x} \neq 0)$ is the solution $\boldsymbol{z}$ over the reals of an equation

$$U\boldsymbol{z} = \boldsymbol{b} \text{ where } \boldsymbol{b} \sim \frac{n}{2}\mathbf{1}, \tag{19}$$

assuming that $B = (B_1, ..., B_k)$ is simple. To prove that $|I_{\boldsymbol{x}}| \sim n/2^k$, we prove the properties of U listed in Lemma 9 below.

Equation (18) implies that by arranging the rows and column indices of U in the same order, U will be symmetric. We will choose an ordering such the first k rows correspond to $B_i, i = 1, ..., k$. Thus $x_i = e_i, i = 1, 2, \ldots, k$ where $e_1 = (1, 0, \ldots, 0)$ etc., and $y_i = e_i, i = 1, 2, \ldots, k$. After this we let Q be the $k \times K$ matrix with column indices x made up of the first k rows.

Thus row i represents $B_i, i = 1, ..., k$ and U contains a $k \times k$ identity matrix in the first k rows and columns.

The row indexed by $\boldsymbol{x} = (x_1, ..., x_k)$ is the linear combination $\sum_{i=1}^k x_i \boldsymbol{r}_i$ of the rows of Q, and corresponds to $B(\boldsymbol{x})$ in the vector space $V(M)$ given above.

Lemma 9. *The $K \times K$ matrix U has the following properties:*

(i) The matrix U symmetric.

(ii) Every row or column of U has 2^{k-1} non-zero entries.

(iii) Any two distinct rows of U have 2^{k-2} common non-zero entries.

(iv) The matrix U is non-singular when the entries are taken to be over the real numbers, and the matrix $S = UU^\top = U^2 = 2^{k-2}(I + J)$ is symmetric, with inverse $S^{-1} = (1/2^{k-2})(I - J/2^k)$; where J is the all-ones matrix.

Proof. (i) This follows immediately from (18), and the above construction.

(ii) Fix $\boldsymbol{x}$ and assume that $x_1 = 1$. There are 2^{k-1} choices for the values of $y_i, i = 2, 3, \ldots, k$. Having made such a choice, there are two choices for y_1, exactly one of which will give $\sum_{i=1}^k x_i y_i = 1$.

(iii) Fix $\boldsymbol{x}, \boldsymbol{x}'$ and think of rows $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{x} + \boldsymbol{x}'$ as non-empty subsets of $[2^k]$. Then we have $|\boldsymbol{x}| = |\boldsymbol{x}'| = |\boldsymbol{x} \setminus \boldsymbol{x}'| + |\boldsymbol{x}' \setminus \boldsymbol{x}| = 2^{k-1}$, by (iii). Thus $|\boldsymbol{x}| + |\boldsymbol{x}'| - (|\boldsymbol{x} \setminus \boldsymbol{x}'| + |\boldsymbol{x}' \setminus \boldsymbol{x}|) = 2|\boldsymbol{x} \cap \boldsymbol{x}'| = 2^{k-1}$.

(iv) That the matrix U is non-singular over the real numbers, uses an argument given in [3] (pages 11-13). Let $S = UU^\top$. Let $\boldsymbol{u}, \boldsymbol{v}$ be distinct rows of U, then $\boldsymbol{u} \cdot \boldsymbol{u} = 2^{k-1}$ and $\boldsymbol{u} \cdot \boldsymbol{v} = 2^{k-2}$. Thus $S = 2^{k-2}(I + J)$, where J is the all-ones matrix. The reader can check that $S^{-1} = \frac{1}{2^{k-2}}(I - \frac{1}{2^k} J)\ 2^{k-1}$ which implies that U is invertible too. □

The definition of a simple k-tuple $(B_1, ..., B_k)$ requires that all sets B_i be large and their set differences to be distinct and of size $\sim n/2$. Let $(|B_1|, \ldots, |B_k|) \sim (n/2)\mathbf{1}$ be the vector of these set sizes. Over the reals, solving (19) gives the sizes of the subsets $I_{\boldsymbol{x}}$.

Lemma 10. *Let $(B_1, ..., B_k)$ be a simple sequence. Then for all $\boldsymbol{x} \in \{0, 1\}^k$,*

$$|I_{\boldsymbol{x}}| = \frac{n}{2^k}\left(1 \pm 2^k \sqrt{\frac{\log n}{n}}\right). \tag{20}$$

Proof. Let $i = 1, ..., K$ index the rows of U and let $B(\boldsymbol{x})$ be the set corresponding to the row $\boldsymbol{x}$ of U. Let $U\boldsymbol{x} = \boldsymbol{b}$ where $b_{\boldsymbol{x}} = 2|B(\boldsymbol{x})|/n = 1 + \varepsilon_i$, where $|\varepsilon_i| \leq 2\sqrt{\log n/n}$. The matrix

$S = U^2$, so $S\boldsymbol{x} = U\boldsymbol{b} = \boldsymbol{c}$ where $c_i = 2^{k-1}(1+\delta_x)$ where $\delta_{\boldsymbol{x}} = \sum \varepsilon_j/2^{k-1}$, the summation being over a 2^{k-1}-subset of rows $\boldsymbol{x}$ of U. Thus, as J is $K \times K$ where $K = 2^k - 1$,

$$\boldsymbol{x} = S^{-1}\boldsymbol{c} = \frac{1}{2^{k-2}}\left(I - \frac{1}{2^k}J\right)\, 2^{k-1}(\mathbf{1}+\boldsymbol{\delta}) = \frac{1}{2^{k-1}}\mathbf{1} + \boldsymbol{\eta},$$

where $|\boldsymbol{\eta}| \leq 2^k\sqrt{\log n/n}$. It follows that w.h.p. the solution to (19) satisfies $|I_{\boldsymbol{x}}| = (n/2^k)(1 \pm 2^k\sqrt{\log n/n})$ for all $\boldsymbol{x} \in \{0,1\}^k$. □

Remark 11. *The proofs above generalize to the case where $\boldsymbol{b} \sim (\xi n, \xi n, \ldots, \xi n)$ for some constant $\xi \in (0, 1/2]$ in equation* (19)*. In which case* (20) *becomes*

$$|I_{\boldsymbol{x}}| = \frac{2\xi n}{2^k}\left(1 \pm 2^k\sqrt{\frac{\log n}{n}}\right).$$

5 Simple sequences of large zero-sum sets.

Let $B_1, B_2, \ldots, B_k$ be a simple sequence. In row M_i of the matrix M, there is a 1 in the diagonal entry $M_{i,i}$. As $s = 3$ there need to be two (random) 1's in column C_i chosen in a way to ensure the linear dependence of $B_1, \ldots, B_k$. The following lemma describes where these non-zeros must be placed.

Lemma 12. *$B_1, \cdots, B_k$ are dependencies if and only if the following holds for all $i \in [n]$. Suppose that row i is in $I_{\boldsymbol{x}}$, and that the two random non-zeros $e_1(i), e_2(i)$ in column i are in $I_{\boldsymbol{u}}, I_{\boldsymbol{v}}$ respectively. Then we must have $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v} (\mathrm{mod}\ 2)$.*

Proof. Let $\boldsymbol{x} = (x_1, ..., x_k)$ and consider x_m for $1 \leq m \leq k$. If $x_m = 0$ then $i \notin B_m$, so either none or both of j, j' are in B_m, and so zero or two unit entries in this column are in B_m. We must therefore have either $u_m = v_m = 0$ or $u_m = v_m = 1$ and $x_m = u_m + v_m$. If $x_m = 1$ then $i \in B_m$ and so exactly one of $e_1(i), e_2(i)$ must also be in B_m. Hence $u_m = 1, v_m = 0$, or vice versa. Thus in all cases $x_m = u_m + v_m$. □

The main result of this section is the following.

Lemma 13. *Let $k \geq 1$ be a positive integer, and let $\mathbf{X}_k$ count the number of simple k-sequences of large dependencies. Then $\mathbf{E}(\mathbf{X}_k) \sim 1$.*

Proof. We have to estimate the expected number of simple sequences $(B_1, ..., B_k)$ of large dependencies. By (20) of Lemma 10 the index sets $I_{\boldsymbol{x}}$ have size $|I_{\boldsymbol{x}}| = (n/2^k)(1+O(\sqrt{\log n/n}))$.

Let $K = 2^k - 1$ as above, and let

$$\Omega = \left\{ \boldsymbol{h} = (h_0, h_1, ..., h_K) : h_i \text{ satisfies (20)}, \sum_{i=1}^{K} h_i \in J_1 \right\}.$$

Then we define $\Phi(\boldsymbol{h}, k)$ by

$$\begin{aligned} \mathbf{E}\left(\mathbf{X}_k\right) =& \sum_{\boldsymbol{h}\in\Omega} \binom{n}{h_0, h_1, \ldots, h_K} \prod_{\boldsymbol{x}\neq 0} \left(2 \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} \frac{h_{\boldsymbol{u}}}{n} \frac{h_{\boldsymbol{v}}}{n} \right)^{h_{\boldsymbol{x}}} \left(\sum_{\boldsymbol{u}} \left(\frac{h_{\boldsymbol{u}}}{n} \right)^2 \right)^{h_0} & (21) \\ =& \sum_{\boldsymbol{h}\in\Omega} \binom{n}{h_0, h_1, \ldots, h_K} \Phi(\boldsymbol{h}, k). & (22) \end{aligned}$$

Explanation of (21). Let $h_{\boldsymbol{x}} = |I_{\boldsymbol{x}}|$. The multinomial coefficient $\binom{n}{h_0,h_1,\ldots,h_K}$ counts the number of choices for the subsets $I_{\boldsymbol{x}}$. In the product, in order for $B_1, ..., B_k$ to be zero-sum, for $\boldsymbol{x} \neq 0$ we need to cancel the diagonal entries $M_{j,j} = 1$ of $j \in I_x$ within the columns indexed by $I_{\boldsymbol{x}}$. This is achieved by putting one entry in rows $I_{\boldsymbol{u}}$ and one in rows $I_{\boldsymbol{v}}$ where $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{x}$. The last factor counts the choices for the entries of columns indexed by I_0 over the row index sets $I_{\boldsymbol{u}}$, either zero or two in an index set, in order to preserve the zero-sum property.

Set $h_{\boldsymbol{x}} = (n/2^k)(1 + \varepsilon_{\boldsymbol{x}})$ where $|\varepsilon_{\boldsymbol{x}}| = O(\sqrt{\log n/n})$. We note that $\sum_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}} = 0$, implies that

$$\sum_{\boldsymbol{x}} h_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}} = \frac{n}{2^k} \sum_{\boldsymbol{x}} (\varepsilon_{\boldsymbol{x}} + \varepsilon_{\boldsymbol{x}}^2) = \frac{n}{2^k} \sum_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}}^2 \text{ and } \sum_{\boldsymbol{x}} h_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}}^2 = \frac{n}{2^k} \sum_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}}^2 + O\left(\frac{\log^{3/2} n}{n^{1/2}} \right).$$

And then Stirling's approximation implies that

$$\begin{aligned} \binom{n}{h_0, h_1, \ldots, h_K} &\sim \frac{n^n \sqrt{2\pi n}}{\prod_{\boldsymbol{x}\in\{0,1\}^k} ((n/2^k)(1 + \varepsilon_{\boldsymbol{x}}))^{h_{\boldsymbol{x}}} (\sqrt{2\pi n/2^k})^{2^k}} \\ &= 2^{kn} \exp\left\{ - \sum_{\boldsymbol{x}\in\{0,1\}^k}^{K} h_{\boldsymbol{x}} \left(\varepsilon_{\boldsymbol{x}} - \frac{\varepsilon_{\boldsymbol{x}}^2}{2} \right) + O(\log n) \right\} \\ &= 2^{kn} \exp\left\{ - \frac{n}{2^{k+1}} \sum_{\boldsymbol{x}\in\{0,1\}^k}^{K} \varepsilon_{\boldsymbol{x}}^2 + O(\log n) \right\} = 2^{kn} n^{O(1)}. \end{aligned}$$

In addition, by considering random 2^k-colorings of $[n]$ we see from the Chernoff bounds that

$$\sum_{\boldsymbol{h}\in\Omega} \binom{n}{h_0, h_1, \ldots, h_K} = 2^{kn}(1 - O(n^{-2^k/3})). \qquad (23)$$

With respect to (21), using $\sum_{\boldsymbol{x}} \varepsilon_{\boldsymbol{x}} = 0$, we see that

$$\begin{aligned}
\left(\sum_{\boldsymbol{u} \in \{0,1\}^k} \left(\frac{h_{\boldsymbol{u}}}{n} \right)^2 \right)^{h_0} &= \left(\sum_{\boldsymbol{u}} \frac{1}{2^{2k}} (1 + 2\varepsilon_{\boldsymbol{u}} + \varepsilon_{\boldsymbol{u}}^2) \right)^{h_0} \\
&= \left(\frac{1}{2^k} \right)^{h_0} \left(1 + \frac{1}{2^k} \sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}}^2 \right)^{h_0} \\
&= \left(\frac{1}{2^k} \right)^{h_0} \exp \left\{ \frac{n}{2^k} (1 + \varepsilon_0) \log \left(1 + \sum_{\boldsymbol{u}} \frac{\varepsilon_{\boldsymbol{u}}^2}{2^k} \right) \right\} \\
&= \left(\frac{1}{2^k} \right)^{h_0} \exp \left\{ \frac{n}{2^{2k}} \sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}}^2 + O\left(\frac{\log^{3/2} n}{n^{1/2}} \right) \right\}. \qquad (24)
\end{aligned}$$

If $\boldsymbol{x} \neq 0$ then each index $\boldsymbol{z}$ occurs exactly once in $\sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} (\varepsilon_{\boldsymbol{u}} + \varepsilon_{\boldsymbol{v}})$ and so $\sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} (\varepsilon_{\boldsymbol{u}} + \varepsilon_{\boldsymbol{v}}) = \sum_{\boldsymbol{z}} \varepsilon_{\boldsymbol{z}} = 0$. Therefore,

$$\begin{aligned}
\left(2 \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} \frac{h_{\boldsymbol{u}}}{n} \frac{h_{\boldsymbol{v}}}{n} \right)^{h_{\boldsymbol{x}}} &= \left(2 \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} \frac{1}{2^{2k}} (1 + \varepsilon_{\boldsymbol{u}} + \varepsilon_{\boldsymbol{v}} + \varepsilon_{\boldsymbol{u}} \varepsilon_{\boldsymbol{v}}) \right)^{h_{\boldsymbol{x}}} \\
&= \left(\frac{1}{2^k} \right)^{h_{\boldsymbol{x}}} \left(1 + \frac{1}{2^k} \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} 2\varepsilon_{\boldsymbol{u}} \varepsilon_{\boldsymbol{v}} \right)^{h_{\boldsymbol{x}}} \\
&= \left(\frac{1}{2^k} \right)^{h_{\boldsymbol{x}}} \exp \left\{ \frac{n}{2^k} (1 + \varepsilon_{\boldsymbol{x}}) \log \left(1 + 2 \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} \frac{\varepsilon_{\boldsymbol{u}} \varepsilon_{\boldsymbol{v}}}{2^k} \right) \right\} \\
&= \left(\frac{1}{2^k} \right)^{h_{\boldsymbol{x}}} \exp \left\{ \frac{n}{2^k} \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} \frac{2\varepsilon_{\boldsymbol{u}} \varepsilon_{\boldsymbol{v}}}{2^k} + O\left(\frac{\log^{3/2} n}{n^{1/2}} \right) \right\}.
\end{aligned}$$

Note that

$$\Lambda = \sum_{\boldsymbol{x} \neq 0} \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} 2\varepsilon_{\boldsymbol{u}} \varepsilon_{\boldsymbol{v}} = \sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}} \sum_{\substack{\boldsymbol{x}+\boldsymbol{u} \\ \boldsymbol{x} \neq 0}} \varepsilon_{\boldsymbol{x}+\boldsymbol{u}} = \sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}} \sum_{\boldsymbol{v} \neq \boldsymbol{u}} \varepsilon_{\boldsymbol{v}},$$

gives

$$\Lambda + \sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}}^2 = \left(\sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}} \right)^2 = 0.$$

Thus using $\sum_{\boldsymbol{x}} h_{\boldsymbol{x}} = n$,

$$\Phi(\boldsymbol{h}, k) = \left(\frac{1}{2^k}\right)^{\sum_{\boldsymbol{x}} h_{\boldsymbol{x}}} \exp\left\{\frac{n}{2^{2k}}\left(\sum_{\boldsymbol{u}} \varepsilon_{\boldsymbol{u}}^2 + \sum_{\boldsymbol{x}\neq 0} \sum_{\substack{\{\boldsymbol{u},\boldsymbol{v}\} \\ \boldsymbol{u}+\boldsymbol{v}=\boldsymbol{x}}} 2\varepsilon_{\boldsymbol{u}}\varepsilon_{\boldsymbol{v}}\right) + O\left(\frac{\log^{3/2} n}{n^{1/2}}\right)\right\}$$
$$= \frac{1}{2^{kn}} e^{O(\log^{3/2} n/\sqrt{n})}. \tag{25}$$

It follows from (22), (23) and (25) above that

$$\mathbf{E}\left(\mathbf{X}_k\right) = 1 + O\left(\frac{\log^{3/2} n}{\sqrt{n}}\right) = 1 + o(1). \tag{26}$$

□

6 Conditional expected number of small zero-sum sets

Let $(B_1, \ldots, B_k)$ be a fixed sequence of subsets of $[n]$ with $|B_i| \in J_1$ for $i = 1, 2, \ldots, k \leq \omega$. Let $\mathcal{B}$ be the event

$$\mathcal{B} = \{(B_1, ...B_k) \text{ is a simple sequence of large row dependencies}\}. \tag{27}$$

We need to understand the conditioning imposed by this event $\mathcal{B}$. Suppose that $|I_{\boldsymbol{x}}| = h_{\boldsymbol{x}} \sim n/2^k$ for $\boldsymbol{x} \in \{0,1\}^k$.

Lemma 14. *Given $\mathcal{B}$ and $i \in I_{\boldsymbol{x}}$, the distribution of the row indices k, ℓ of the other two non-zeros in column i is as follows: if $\boldsymbol{x} \neq 0$ then choose $\boldsymbol{u}, \boldsymbol{v}$ such that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v} \bmod 2$ with probability*

$$p(\boldsymbol{u}, \boldsymbol{v}) = \frac{2h_{\boldsymbol{u}}h_{\boldsymbol{v}}}{\sum_{\boldsymbol{y}+\boldsymbol{z}=\boldsymbol{x}} h_{\boldsymbol{y}}h_{\boldsymbol{z}}},$$

and then randomly choose $k \in I_{\boldsymbol{u}}, \ell \in I_{\boldsymbol{v}}$. If $\boldsymbol{x} = 0$ then choose $\boldsymbol{u}$ with probability

$$p(\boldsymbol{u}, \boldsymbol{u}) = \frac{h_{\boldsymbol{u}}^2}{\sum_{\boldsymbol{y}\in\{0,1\}^k} h_{\boldsymbol{y}}^2},$$

and then randomly choose $k, \ell \in I_{\boldsymbol{u}}$.

Proof. This follows from the fact that the non-zeros in each column are independently chosen with replacement and from the condition given in Lemma 12. □

Let $(S_j,\ s_j = |S_j| \leq \omega,\ j = 1, 2, \ldots, \ell \leq \omega)$ be a sequence of pairwise disjoint small subsets of $[n]$ and $S = \bigcup_{j=1}^{\ell} S_j$ and $s = |S|$. We define the events

$$\mathcal{S}_j = \{S_j \text{ is a small zero-sum row set}\} \text{ for } j = 1, 2, \ldots, \ell \text{ and } \mathcal{S} = \bigcap_{j=1}^{\ell} \mathcal{S}_j.$$

$$\mathcal{S}_j^* = \{S_j \text{ is a small fundamental zero-sum row set}\} \text{ for } j = 1, 2, \ldots, \ell \text{ and } \mathcal{S}^* = \bigcap_{j=1}^{\ell} \mathcal{S}_j^*.$$

Lemma 15.

$$\mathbb{P}(\mathcal{S}^* \mid \mathcal{B}) \sim \mathbb{P}(\mathcal{S}^*). \tag{28}$$

Proof. Let $I_{\boldsymbol{x}}$, $\boldsymbol{x} \in \{0,1\}^k$, be as defined in Section 4. Let $S_{\boldsymbol{x}} = S \cap I_{\boldsymbol{x}}$ and $J_{j,\boldsymbol{x}} = S_j \cap I_{\boldsymbol{x}}$ and $\ell_{j,\boldsymbol{x}} = |J_{j,\boldsymbol{x}}|$ for $i = 1, 2, \ldots, m$ and $J_{\boldsymbol{x}} = \bigcup_{j=1}^{m} J_{j,\boldsymbol{x}}$ and $\ell_{\boldsymbol{x}} = |J_{\boldsymbol{x}}|$. Let $J_{\mathbf{0}} = I_{\mathbf{0}} \setminus S$ and $\ell_{\mathbf{0}} = |S_{\mathbf{0}}|$. We now consider the probability that column i is consistent with $\mathcal{S}$. We let $h_{\boldsymbol{x}} = |I_{\boldsymbol{x}}|$ and $s_{\boldsymbol{x}} = |S_{\boldsymbol{x}}|$ for $\boldsymbol{x} \in \{0,1\}^k$.

Case 1: $i \in I_{\mathbf{0}} \setminus J_{\mathbf{0}}$**.** For each column $i \in I_{\mathbf{0}} \setminus J_{\mathbf{0}}$, the task here is to estimate the probability that the two non-zeros $e_1(i), e_2(i)$ are in rows consistent with the occurrence of $\mathcal{S}$. Because $i \in I_{\mathbf{0}}$ and $\mathcal{B}$ occurs, we know from Lemma 12 that $e_1(i), e_2(i) \in I_{\boldsymbol{u}}$ for some $\boldsymbol{u} \in \{0,1\}^k$. For $\mathcal{S}$ to occur, we require that zero or two of $e_1(i), e_2(i)$ fall in $J_{\boldsymbol{u}}$, an event of conditional probability $(1 - s_{\boldsymbol{u}}/h_{\boldsymbol{u}})^2 + (s_{\boldsymbol{u}}/h_{\boldsymbol{u}})^2$.

Let $E_{\boldsymbol{u}}$ denote the number of non-zero pairs from $I_{\mathbf{0}} \setminus J_{\mathbf{0}}$ falling in $J_{\boldsymbol{u}}$. Then the conditional probability that the non-zeros of $I_{\mathbf{0}} \setminus S_{\mathbf{0}}$ are consistent with $\mathcal{S}$ is given by

$$\mathbb{P}(I_{\mathbf{0}} \setminus S_{\mathbf{0}} \text{ is consistent } \mathcal{S} \mid \mathcal{B}) = \mathbf{E}\left(\prod_{\boldsymbol{u}} \left(1 - 2\frac{s_{\boldsymbol{u}}}{h_{\boldsymbol{u}}} + 2\left(\frac{s_{\boldsymbol{u}}}{h_{\boldsymbol{u}}}\right)^2\right)^{E_{\boldsymbol{u}}}\right) \tag{29}$$

Given $\mathcal{B}$, we see that $E_{\boldsymbol{u}}$ is distributed as $Bin(h_{\mathbf{0}} - s_{\mathbf{0}}, p(\boldsymbol{u}, \boldsymbol{u}))$, and has expectation

$$\mathbf{E}\left(E_{\boldsymbol{u}}\right) = (h_{\mathbf{0}} - s_{\mathbf{0}})\frac{h_{\boldsymbol{u}}^2}{h_{\mathbf{0}}^2 + h_1^2 + \cdots + (h_{2^k-1})^2} \sim \frac{h_{\mathbf{0}}}{2^k}.$$

The Chernoff bounds imply that $E_{\boldsymbol{u}}$ is concentrated around its mean $(h_{\mathbf{0}} - s_{\mathbf{0}})p(\boldsymbol{u}, \boldsymbol{u}) \sim \frac{N}{2^k}$, where $N = n/2^k$. Thus,

$$\left|E_{\boldsymbol{u}} - \frac{h_{\mathbf{0}}}{2^k}\right| \leq n^{2/3} \quad \text{with probability at least } 1 - e^{-\Omega(n^{1/3})}. \tag{30}$$

Going back to (29) and using (30) we have

$$\mathbb{P}(I_{\mathbf{0}} \setminus S_{\mathbf{0}} \text{ is consistent with the occurrence of } \mathcal{S} \mid \mathcal{B}) \sim$$
$$\prod_{\boldsymbol{u}} \left(1 - \frac{2s_{\boldsymbol{u}}}{N}\right)^{N/2^k} \sim \exp\left\{-2\sum_{\boldsymbol{u}} \frac{s_{\boldsymbol{u}}}{2^k}\right\} = e^{-s/2^{k-1}}. \quad (31)$$

Case 2: $i \in I_{\boldsymbol{x}} \setminus J_{\boldsymbol{x}},\ \boldsymbol{x} \neq \mathbf{0}$. Given $\mathcal{B}$, and $i \in I_{\boldsymbol{x}}$, suppose that the non-zeros $e_1(i), e_2(i)$ of column i lie in $I_{\boldsymbol{u}}, I_{\boldsymbol{x}+\boldsymbol{u}}$ respectively, $\boldsymbol{u} \in \{0,1\}^k$. The probability of this is $p(\boldsymbol{u}, \boldsymbol{x}+\boldsymbol{u})$. The number $E_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{x}+\boldsymbol{u})$ of such pairs of non-zeros in $I_{\boldsymbol{u}}, I_{\boldsymbol{x}+\boldsymbol{u}}$ has distribution $Bin((h_{\boldsymbol{x}} - s_{\boldsymbol{x}})p(\boldsymbol{u}, \boldsymbol{x}+\boldsymbol{u}))$, and expectation asymptotic to $(h_{\boldsymbol{x}} - s_{\boldsymbol{x}})/2^{k-1}$.

The rows of $S_1, \ldots, S_\ell$ have to be zero-sum in this column, so either exactly one non-zero falls in each of $S_{j,\boldsymbol{u}}, S_{j,\boldsymbol{x}+\boldsymbol{u}}$ for some $1 \leq j \leq \ell$ or exactly one non-zero falls in each of $I_{\boldsymbol{u}} \setminus S_{\boldsymbol{u}}, I_{\boldsymbol{x}+\boldsymbol{u}} \setminus S_{\boldsymbol{x}+\boldsymbol{u}}$. The probability of this is

$$\begin{aligned} P(\boldsymbol{u}, \boldsymbol{x}+\boldsymbol{u}) =& \mathbf{E}\left(\left(\sum_{j=1}^{\ell} \frac{s_{j,\boldsymbol{u}}}{h_{\boldsymbol{u}}}\frac{s_{j,\boldsymbol{x}+\boldsymbol{u}}}{h_{\boldsymbol{x}+\boldsymbol{u}}} + \frac{h_{\boldsymbol{u}} - s_{\boldsymbol{u}}}{h_{\boldsymbol{u}}}\frac{h_{\boldsymbol{x}+\boldsymbol{u}} - s_{\boldsymbol{x}+\boldsymbol{u}}}{h_{\boldsymbol{x}+\boldsymbol{u}}}\right)^{E_{\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{x}+\boldsymbol{u})}\right) \\ &\sim \left(\sum_{j=1}^{\ell} \frac{s_{j,\boldsymbol{u}}s_{j,\boldsymbol{x}+\boldsymbol{u}}}{N^2} + \frac{N - s_{\boldsymbol{u}}}{N}\frac{N - s_{\boldsymbol{x}+\boldsymbol{u}}}{N}\right)^{(N-s_{\boldsymbol{x}})/2^{k-1}} \\ &\sim e^{-(s_{\boldsymbol{u}}+s_{\boldsymbol{x}+\boldsymbol{u}})/2^{k-1}}. \end{aligned}$$

For a given $\boldsymbol{x}$ there are 2^{k-1} unordered pairs $S_{\boldsymbol{u}}, S_{\boldsymbol{x}+\boldsymbol{u}}$, so

$$\mathbb{P}(I_{\boldsymbol{x}} \setminus S_{\boldsymbol{x}} \text{ is consistent with } \mathcal{S}) \sim \exp\left\{-\frac{1}{2^{k-1}}\sum_{\{u,\boldsymbol{x}+\boldsymbol{u}\}}(s_{\boldsymbol{u}} + s_{\boldsymbol{x}+\boldsymbol{u}})\right\} = e^{-s/2^{k-1}}. \quad (32)$$

(In the sum in (32) $s_{\boldsymbol{u}} + s_{\boldsymbol{x}+\boldsymbol{u}}$ and $s_{\boldsymbol{x}+\boldsymbol{u}} + s_{\boldsymbol{u}}$ contribute as one term.) Thus

$$\mathbb{P}(I_{\boldsymbol{x}} \setminus S_{\boldsymbol{x}} \text{ is consistent with } \mathcal{S}, \forall \boldsymbol{x} \neq \mathbf{0}) \sim e^{-(2^k-1)s/2^{k-1}}. \quad (33)$$

Case 3: $i \in S_{j,\boldsymbol{x}} \subseteq I_{\boldsymbol{x}},\ \boldsymbol{x} \neq 0$. For $i \in S_{j,\boldsymbol{x}}$, one non-zero needs to be in S_j, and the other to avoid S_j. Let $\boldsymbol{v} = \boldsymbol{x} + \boldsymbol{u}$. Suppose that the pair $e_1(i)$, $e_2(i)$ fall in $I_{\boldsymbol{u}}, I_{\boldsymbol{u}+\boldsymbol{x}}$. The probability this happens is

$$P_j(\boldsymbol{u}, \boldsymbol{v}) \sim \frac{1}{2^{k-1}}\left(\frac{s_{j,\boldsymbol{u}}}{h_{\boldsymbol{u}}}\frac{h_{\boldsymbol{v}} - s_{j,\boldsymbol{v}}}{h_{\boldsymbol{v}}} + \frac{s_{j,\boldsymbol{v}}}{h_{\boldsymbol{v}}}\frac{h_{\boldsymbol{u}} - s_{j,\boldsymbol{u}}}{h_{\boldsymbol{u}}}\right). \quad (34)$$

The events $\{\boldsymbol{u}, \boldsymbol{x}+\boldsymbol{u}\}$ are disjoint and exhaustive, so for a given $i \in S_{j,\boldsymbol{x}}$ the probability $p(i,j)$ of success (i.e. the S_j-indexed rows of column i sum to zero) is

$$p(i,j) = \sum_{\{\boldsymbol{u},\boldsymbol{u}+\boldsymbol{x}\}} P_j(\boldsymbol{u}, \boldsymbol{u}+\boldsymbol{x}) \sim \frac{1}{2^{k-1}} \sum_{\boldsymbol{u},\boldsymbol{v}=\boldsymbol{x}+\boldsymbol{u}} \left(\frac{s_{j,\boldsymbol{u}}}{N} \frac{N - s_{j,\boldsymbol{v}}}{N} + \frac{s_{j,\boldsymbol{v}}}{N} \frac{N - s_{j,\boldsymbol{u}}}{N} \right)$$
$$\sim \frac{s_j}{N 2^{k-1}} \left(1 + O\left(\frac{\omega}{N} \right) \right). \quad (35)$$

Every column of $S_{j,\boldsymbol{x}}$ has to succeed or S_j is not a small zero-sum set. Thus

$$\mathbb{P}(S_{\boldsymbol{x}} \text{ succeeds}) \sim \left(\frac{s_j(1 + O(s/N))}{N 2^{k-1}} \right)^{s_{j,\boldsymbol{x}}}.$$

As $\sum s_{j,\boldsymbol{x}} = s_j$,

$$\mathbb{P}(S_{\boldsymbol{x}} \text{ succeeds } \forall \boldsymbol{x}) \sim \left(\frac{s_j}{N 2^{k-1}} \right)^{s_j - s_{j,\boldsymbol{0}}}. \quad (36)$$

Case 4: $i \in S_{j,\boldsymbol{0}} \subseteq I_{j,\boldsymbol{0}}$**.** In the case that $\boldsymbol{x} = \boldsymbol{0}$, and $S_{j,\boldsymbol{0}} \subseteq I_{j,\boldsymbol{0}}$, the non-zeros in a column of $S_{j,\boldsymbol{0}}$ must both fall in the same index set $I_{\boldsymbol{u}}$; one in $S_{j,\boldsymbol{u}}$ and one in $I_{\boldsymbol{u}} \setminus S_{j,\boldsymbol{u}}$. Thus $P(\boldsymbol{u}, \boldsymbol{u})$ is now summed over all $I_{\boldsymbol{u}}$, a total of 2^k such sets. For $i \in S_{j,\boldsymbol{0}}$, the probability $p(i)$ of success is

$$p(i) = \sum_{\{\boldsymbol{u},\boldsymbol{u}\}} P(\boldsymbol{u}, \boldsymbol{u}) \sim \frac{1}{2^k} \sum_{\boldsymbol{u}} \left(2 \frac{s_{j,\boldsymbol{u}}}{N} \frac{N - s_{j,\boldsymbol{u}}}{N} \right) \sim \frac{s_j}{N 2^{k-1}} \left(1 + O\left(\frac{\omega}{N} \right) \right).$$

The final term is the same as in (35), and we obtain

$$\mathbb{P}(S_{j,\boldsymbol{0}} \text{ succeeds}) \sim \left(\frac{s_j}{N 2^{k-1}} \right)^{s_{j,\boldsymbol{0}}} \quad (37)$$

Using (31), (33), (36) and (37), we obtain

$$\mathbb{P}(\mathcal{S} \mid \mathcal{B}) \sim \prod_{j=1}^{m} \left(\frac{s_j}{N 2^{k-1}} \right)^{s_j} e^{-(2^k-1)s/2^{k-1}} e^{-s/2^{k-1}} = \prod_{j=1}^{m} \left(\frac{2 s_j}{n} \right)^{s_j} e^{-2s}, \quad (38)$$

after using (5). This completes the proof of $\mathbb{P}(\mathcal{S} \mid \mathcal{B}) \sim \mathbb{P}(\mathcal{S})$. To replace $\mathcal{S}$ by $\mathcal{S}^*$ we just need to let $K_j, j = 1, 2, \ldots, m$ denote the set of i in Case 3 where $i \in S_{j,\boldsymbol{x}}$. We see from (34) that the positions of the non-zeros in the columns K_j are asymptotically uniform over S_j. This is because each $k \in J_{j,\boldsymbol{u}}$ is chosen with probability asymptotic to $\frac{1}{s_{j,\boldsymbol{u}}} \cdot \frac{s_{j,\boldsymbol{u}}}{h_{\boldsymbol{u}}}$ and similarly for $k \in J_{j,\boldsymbol{x}+\boldsymbol{u}}$. In which case, the conditional probability that S_j is fundamental is obtained by multiplying by κ_{s_j}. This completes the proof of the lemma. $\square$

We can now use inclusion-exclusion to prove

Lemma 16. *Let Σ_σ be the event that there are exactly σ disjoint small fundamental dependencies. Then,*

$$\mathbb{P}(\Sigma_\sigma \mid \mathcal{B}) \sim \frac{\phi_R^\sigma e^{-\phi_R}}{\sigma!} \sim \mathbb{P}(\Sigma_\sigma).$$

Proof. Let

$$T_\ell = \frac{1}{\ell!} \sum_{1 \le s_1,\ldots,s_\ell \le \omega} \sum_{|S_i|=s_i, i=1,\ldots,\ell} \mathbb{P}\left(\bigcap_{i=1}^{\ell} \mathcal{S}_i^* \middle| \mathcal{B} \right) \sim \frac{1}{\ell!} \sum_{1 \le s_1,\ldots,s_\ell \le \omega} \sum_{|S_i|=s_i, i=1,\ldots,\ell} \mathbb{P}\left(\bigcap_{i=1}^{\ell} \mathcal{S}_i^* \right) \sim$$
$$\frac{1}{\ell!} \sum_{1 \le s_1,\ldots,s_\ell \le \omega} \binom{n}{s_1,\ldots,s_\ell} \prod_{i=1}^{\ell} \left(\frac{2s_i}{n} \right)^{s_i} e^{-2s_i} \kappa_{s_i} \sim \frac{1}{\ell!} \sum_{1 \le s_1,\ldots,s_\ell \le \omega} \prod_{i=1}^{\ell} \frac{(2s_i)^{s_i}}{s_i!} e^{-2s_i} \kappa_{s_i}$$
$$\sim \frac{1}{\ell!} \left(\sum_{s=1}^{\infty} \frac{(2e^{-2})^s}{s} \sigma_s \right)^{\ell} \sim \frac{\phi_R^\ell}{\ell!}.$$

The first approximation follows from Lemma 15 and the second from (5), (6).

Using Inclusion-Exclusion, we have

$$\mathbb{P}(\Sigma_\sigma \mid \mathcal{B}) = \sum_{\ell \ge \sigma} (-1)^{k-\sigma} \binom{\ell}{\sigma} T_\ell \sim \sum_{\ell \ge \sigma} (-1)^{\ell-\sigma} \binom{\ell}{\sigma} \frac{\phi_R^\ell}{\ell!} = \frac{\phi_R^\sigma e^{-\phi_R}}{\sigma!}.$$

Lemma 8 gives us the unconditional probability. □

Let $\mathbf{X}_k$ count the number of simple k-sequences as in Lemma 13.

Lemma 17. *If $\sigma = O(1)$ then $\mathbf{E}(\mathbf{X}_k \mid \Sigma_\sigma) \sim 1$.*

Proof.

$$
\begin{aligned}
\mathbf{E}\left(\mathbf{X}_k \mid \Sigma_\sigma\right) &= \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \mathbb{P}(\mathcal{B} \mid \Sigma_\sigma) \\
&= \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \frac{\mathbb{P}(\Sigma_\sigma \mid \mathcal{B})\mathbb{P}(\mathcal{B})}{\mathbb{P}(\Sigma_\sigma)} \\
&= \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \frac{\mathbb{P}(\mathcal{B})}{\mathbb{P}(\Sigma_\sigma)} \sum_{\ell\geq\sigma}(-1)^{\ell-\sigma}\binom{k}{\sigma}T_\ell \\
&= \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \frac{\mathbb{P}(\mathcal{B})}{\mathbb{P}(\Sigma_\sigma)} \sum_{\ell\geq\sigma}(-1)^{\ell-\sigma}\binom{k}{\sigma}\frac{1}{\ell!} \sum_{1\leq s_1,\ldots,s_\ell\leq\omega} \sum_{|S_i|=s_i, i=1,\ldots,\ell} \mathbb{P}\left(\bigcap_{i=1}^{\ell} \mathcal{S}_i^* \middle| \mathcal{B}\right) \\
&\sim \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \frac{\mathbb{P}(\mathcal{B})}{\mathbb{P}(\Sigma_\sigma)} \sum_{\ell\geq\sigma}(-1)^{\ell-\sigma}\binom{k}{\sigma}\frac{1}{\ell!} \sum_{1\leq s_1,\ldots,s_\ell\leq\omega} \sum_{|S_i|=s_i, i=1,\ldots,\ell} \mathbb{P}\left(\bigcap_{i=1}^{\ell} \mathcal{S}_i^*\right) \\
&\sim \sum_{\mathcal{B}=(B_1,\ldots,B_k)} \frac{\mathbb{P}(\mathcal{B})}{\mathbb{P}(\Sigma_\sigma)}\mathbb{P}(\Sigma_\sigma) \\
&= \mathbf{E}\left(\mathbf{X}_k\right) \sim 1.
\end{aligned}
$$

□

7 Joint distribution of small and large dependencies

We first state a preparatory lemma. Let π be the probability distribution given by

$$
\pi(k) = \begin{cases} \prod_{j=1}^{\infty}\left(1-\left(\frac{1}{2}\right)^j\right) & k = 0. \\ \frac{\prod_{j=k+1}^{\infty}\left(1-\left(\frac{1}{2}\right)^j\right)}{\prod_{j=1}^{k}\left(1-\left(\frac{1}{2}\right)^j\right)}\left(\frac{1}{2}\right)^{k^2} & k \geq 1. \end{cases} \tag{39}
$$

A proof of the next result for $c_k = 1$ can be found in [6], [7]. We give a full and different proof for completeness.

Lemma 18. *For $\lambda \geq 0$, the solutions to*

$$
c_k = \sum_{\lambda=k}^{\infty} q_\lambda \prod_{i=0}^{k-1}(2^\lambda - 2^i), \qquad k \geq 0. \tag{40}
$$

are given by

$$q_\lambda = \sum_{k=\lambda}^{\infty} (-1)^{k-\lambda} 2^{\binom{k-\lambda}{2}} \begin{bmatrix} k \\ \lambda \end{bmatrix}_2 \psi_k c_k, \tag{41}$$

where $\psi_k = 1/\left(2^{\binom{k}{2}} \prod_{i=1}^{k} (2^i - 1)\right)$. *In particular, if* $c_k = 1$, $q_\lambda = \pi(\lambda)$ *of* (39).

Proof. Gaussian coefficients are defined as

$$\begin{bmatrix} \lambda \\ k \end{bmatrix}_z = \frac{\prod_{i=1}^{k} (z^{\lambda-i+1} - 1)}{\prod_{i=1}^{k} (z^i - 1)}. \tag{42}$$

Using (42) with $z = 2$, equation (40) can be rewritten as

$$c_k = 2^{\binom{k}{2}} \prod_{i=1}^{k} (2^i - 1) \sum_{\lambda=k}^{\infty} q_\lambda \begin{bmatrix} \lambda \\ k \end{bmatrix}_2. \tag{43}$$

Put $\psi_k = 1/\left(2^{\binom{k}{2}} \prod_{i=1}^{k} (2^i - 1)\right)$, we see that q_λ is the solution to

$$\sum_{\lambda=k}^{\infty} \begin{bmatrix} \lambda \\ k \end{bmatrix}_2 q_\lambda = \psi_k c_k, \qquad k \geq 0. \tag{44}$$

Fix $\delta \geq 0$, multiply equation $k \geq \delta$ in (44) by $(-1)^{k-\delta} 2^{\binom{k-\delta}{2}} \begin{bmatrix} k \\ \delta \end{bmatrix}_2$, and sum these equations over $k \geq \delta$. This gives

$$\begin{aligned} \sum_{k=\delta}^{\infty} (-1)^{k-\delta} 2^{\binom{k-\delta}{2}} \begin{bmatrix} k \\ \delta \end{bmatrix}_2 \psi_k c_k &= \sum_{k=\delta}^{\infty} \sum_{\lambda=k}^{\infty} (-1)^{k-\delta} \begin{bmatrix} k \\ \delta \end{bmatrix}_2 2^{\binom{k-\delta}{2}} \begin{bmatrix} \lambda \\ k \end{bmatrix}_2 q_\lambda & (45) \\ &= \sum_{k=\delta}^{\infty} \sum_{\lambda=k}^{\infty} (-1)^{k-\delta} \begin{bmatrix} \lambda - \delta \\ k - \delta \end{bmatrix}_2 2^{\binom{k-\delta}{2}} \begin{bmatrix} \lambda \\ \delta \end{bmatrix}_2 q_\lambda & \\ &= \sum_{\lambda=\delta}^{\infty} \begin{bmatrix} \lambda \\ \delta \end{bmatrix}_2 q_\lambda \sum_{k=\delta}^{\lambda} (-1)^{k-\delta} \begin{bmatrix} \lambda - \delta \\ k - \delta \end{bmatrix}_2 2^{\binom{k-\delta}{2}} & (46) \\ &= q_\delta. & (47) \end{aligned}$$

Explanation: (46) **to** (47)**:** Gaussian coefficients satisfy the identity

$$(1 + x)(1 + zx) \cdots (1 + z^{r-1} x) = \sum_{\ell=0}^{r} \begin{bmatrix} r \\ \ell \end{bmatrix}_z z^{\binom{\ell}{2}} x^\ell. \tag{48}$$

To prove the last summation on the right hand side of (46) is zero for $\lambda > \delta$, use (48) with $x = -1, z = 2$, $\ell = k - \delta$ and $r = \lambda - \delta$. This gives $\sum_{\ell=0}^{\lambda-\delta} {\lambda-\delta \brack \ell}_2 2^{\binom{\ell}{2}} (-1)^\ell = 0$ for $\lambda > \delta$.
For $z < 1$, taking the limit of (48) gives

$$\prod_{\ell=0}^{\infty}(1 + z^\ell x) = \sum_{\ell=0}^{\infty} \frac{z^{\binom{\ell}{2}} x^\ell}{\prod_{i=1}^{\ell}(1 - z^i)}. \tag{49}$$

Replacing δ by λ, and putting $c_k = 1$ in (41), we see that the solution q_λ to (40) is

$$\begin{aligned} q_\lambda &= \sum_{k=\lambda}^{\infty} \frac{(-1)^{k-\lambda} 2^{\binom{k-\lambda}{2} - \binom{k}{2}}}{\prod_{i=0}^{\lambda-1}(2^{\lambda-i} - 1) \prod_{i=\lambda}^{k-1}(2^{k-i} - 1)} \\ &= \frac{\left(\frac{1}{2}\right)^{\lambda^2}}{\prod_{i=1}^{\lambda}\left(1 - \left(\frac{1}{2}\right)^i\right)} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell \left(\frac{1}{2}\right)^{\binom{\ell}{2}} \left(\frac{1}{2}\right)^{(1+\lambda)\ell}}{\prod_{i=1}^{\ell}\left(1 - \left(\frac{1}{2}\right)^i\right)} & (50) \\ &= \left(\frac{1}{2}\right)^{\lambda^2} \frac{\prod_{i=\lambda+1}^{\infty}\left(1 - \left(\frac{1}{2}\right)^i\right)}{\prod_{i=1}^{\lambda}\left(1 - \left(\frac{1}{2}\right)^i\right)} = \pi(\lambda), & (51) \end{aligned}$$

where $\pi(\lambda)$ is given in (39). To get from (50) to (51), use (49) with $z = 1/2$ and $x = (-1/2^{\lambda+1})$. $\square$

Quotient space argument

Given M, let $\mathcal{B} = \{B_i : i \in [N]\}$ denote the set of large dependencies and $\mathcal{S} = \{S_j : j \in [T]\}$ denote the set of small dependencies. The following observations complete the proof of Theorem 1.

P1 Suppose that V, V_S are the vector spaces generated by all dependencies, and small dependencies, respectively. Suppose that these spaces have dimensions d, σ respectively.

Let $W = V/V_S$ be the quotient space and f_S be the canonical map $f_S : V \to W$. Thus f_S maps small dependencies to zero and $W = \{f_S(B) : B \in \mathcal{B}\} \cup \{0\}$. Each vector in W corresponds to an equivalence class of vectors in V. In terms of dependencies in $\mathcal{B}$, $B \sim B'$ iff $B \oplus B' = S$ where $S \in \mathcal{S}$. As the small dependencies are disjoint, the size of the equivalence class of B is 2^σ.

P2 Note that $dim(W) = dim(V) - dim(V_S) = d - \sigma$. Let λ denote the maximum number of independent large dependencies. This will be the same as the maximum length of a simple sequence. We next prove that $\lambda = dim(W)$.

Let $\mathbf{b}_i, i = 1, 2, \ldots, m$ be a basis of W then $B_i \in {f_S}^{-1}(\mathbf{b}_i), i = 1, 2, \ldots, m$ form a simple sequence. If not then for some $A \subseteq [m]$ we have $\oplus_{i\in A} B_i \in V_S$ which implies that $f_S\left(\oplus_{i\in A} B_i\right) = \sum_{i\in A} \mathbf{b}_i = 0$. Conversely, if $B_1, B_2, \ldots, B_k$ is a simple sequence then $\mathbf{b}_i = f_S(B_i), i = 1, 2, \ldots, k$ are independent. If not then for some $A \subseteq [k]$, $\sum_{i\in A} \mathbf{b}_i = 0$ which implies that $\oplus_{i\in A} B_i \in V_S$.

P3 The first i independent members of a simple sequence generate a vector space W_i of size 2^i. The next independent entry of the sequence is chosen from $W \setminus W_i$, a space of size $2^\lambda - 2^i$. Each entry is chosen from an equivalence class of size 2^σ. It follows that the number X_k of simple sequences of length k is equal to

$$\prod_{i=0}^{k-1} ((2^\lambda - 2^i) \times 2^\sigma) = 2^{k\sigma} \prod_{i=0}^{k-1} (2^\lambda - 2^i).$$

P4 Let $b_t = \mathbb{P}(\lambda = t \mid \sigma = s)$. By Lemma 17, $\mathbf{E}\left(X_k \mid \sigma = s\right) \sim 1$, so

$$1 \sim \mathbf{E}\left(X_k \mid \sigma = s\right) \;=\; 2^{sk} \sum_{t=k}^{\infty} \prod_{i=0}^{k-1} (2^t - 2^i) b_t. \tag{52}$$

This can be re-written (with $\sim$ replaced by $=$) as,

$$2^{-sk} = 2^{\binom{k}{2}} \prod_{i=1}^{k} (2^i - 1) \; \sum_{t=k}^{\infty} b_t \begin{bmatrix} t \\ k \end{bmatrix}_2$$

By Lemma 18 we find that

$$\begin{aligned}
b_t &= \sum_{k=t}^{\infty} (-1)^{k-t} 2^{\binom{k-t}{2}} \begin{bmatrix} k \\ t \end{bmatrix}_2 \psi_k c_k \\
&= \sum_{k=t}^{\infty} \frac{(-1)^{k-t} 2^{\binom{k-t}{2} - \binom{k}{2} - ks}}{\prod_{i=1}^{k} (2^i - 1)} \begin{bmatrix} k \\ t \end{bmatrix}_2 \\
&= \frac{1}{(2^t - 1) \cdots (2 - 1)} \sum_{k \geq t} (-1)^{k-t} 2^{\binom{k-t}{2} - \binom{k}{2} - ks - \binom{k+1-t}{2}} \frac{1}{\prod_{i=1}^{k-t} (1 - (1/2)^i)} \\
&= \left(\frac{1}{2}\right)^{t(t+s)} \frac{1}{\prod_{j=1}^{t} (1 - 1/2^j)} \sum_{j \geq 0} \left(\frac{1}{2}\right)^{\binom{j}{2}} \left(-1 \left(\frac{1}{2}\right)^{1+s+t}\right)^j \frac{1}{\prod_{i=1}^{j} (1 - (1/2)^i)} \\
&= \left(\frac{1}{2}\right)^{t(t+s)} \frac{1}{\prod_{j=1}^{t} (1 - 1/2^j)} \prod_{j=0}^{\infty} \left(1 - \left(\frac{1}{2}\right)^{(s+t+1)+j}\right) \\
&= P(s, t),
\end{aligned}$$

as given in (2), and where we used (49) with $z = 1/2$ and $x = -(1/2)^{s+t+1}$ to replace the alternating sum.

P5 The $P(s,t)$ only asymptotically satisfy the solution $b_t(s) = \mathbb{P}(\lambda = t \mid \sigma = s)$ in (52) asymptotically. So to prove the lemma, we show that for large K,

$$\sum_{\substack{t \geq K \\ s \geq 0}} b_t(s) \leq \varepsilon, \tag{53}$$

where $\varepsilon > 0$ is arbitrarily small. For $t \geq k$,

$$\prod_{i=0}^{k-1}(2^t - 2^i) = 2^{kt}\prod_{i=0}^{k-1}\left(1 - \frac{1}{2^{t-i}}\right) \geq 2^{kt}\left(1 - \sum_{i=0}^{k-1}\frac{1}{2^{t-i}}\right) \geq 2^{(k-1)t}.$$

It follows that

$$\sum_{\substack{t \geq K \\ s \geq 0}} b_t(s) \leq 2^{-K(K-1)}.$$

Thus (53) holds if $K \geq \sqrt{2\log_2 1/\varepsilon}$.

References

[1] D. Achiloptas and M. Molloy. The solution space geometry of random linear equations, *Random Structures and Algorithms* 46.2, 197–231, (2015).

[2] B. Bollobas. *Random Graphs*, 2nd edition. Cambridge University Press (2001).

[3] R. Brualdi and H. Ryser. *Combinatorial Matrix Theory.* Cambridge University Press. (1991).

[4] T. Bohman and A.M. Frieze. Hamilton cycles in 3-out, *Random Structures and Algorithms* 35, 393-417, (2009).

[5] A. Coja-Oghlan, A. Ergür, P. Gao, S. Hetterich, M. Rolvien. The rank of sparse random matrices, SODA 2020, 579-591, (2020).

[6] C. Cooper. On the rank of random matrices, *Random Structures and Algorithms* 16, 209-232, (2000).

[7] C. Cooper. On the distribution of rank of a random matrix over a finite field, *Random Structures and Algorithms* 17, 197-212, (2000).

[8] C. Cooper, A.M. Frieze and W. Pegden. On the rank of a random binary matrix, SODA 2019, 946-955, (2019).

[9] C. Cooper and A.M. Frieze. Rank of the vertex-edge incidence matrix of r-out hypergraphs. Extended ArXiv version (2021). `https://arxiv.org/pdf/2107.05779.pdf`

[10] T. Fenner and A.M. Frieze. On the connectivity of random m-orientable graphs and digraphs, *Combinatorica* 2, 347-359, (1982).

[11] A.M. Frieze. Maximum matchings in a class of random graphs, *Journal of Combinatorial Theory B* 40, 196-212, (1986).

[12] A.M. Frieze and M. Karoński. *Introduction to Random Graphs*, Cambridge University Press, (2016).

[13] M. Ibrahimi, Y. Kanoria, M. Kraning and A. Montanari. The set of solutions of random XORSAT formulae, *Annals of Applied Probability*, 25.5, 2743–2808, (2015).

[14] I. N. Kovalenko, A. A. Levitskya and M. N. Savchuk. *Selected Problems in Probabilistic Combinatorics.* Naukova Dumka, Kyiv (1986) (in Russian).