SUPERBOOLEAN RANK AND THE SIZE OF THE LARGEST TRIANGULAR SUBMATRIX OF A RANDOM MATRIX

ZUR IZHAKIAN, SVANTE JANSON, AND JOHN RHODES

Abstract. We explore the size of the largest (permuted) triangular submatrix of a random matrix, and more precisely its asymptotical behavior as the size of the ambient matrix tends to infinity. The importance of such permuted triangular submatrices arises when dealing with certain combinatorial algebraic settings in which these submatrices determine the rank of the ambient matrix, and thus attract a special attention.

1. Introduction

Let $X = X_n = (x_{ij})_{i,j=1}^n$ be a random $n \times n$ matrix. We assume that the entries of X_n are taken from some set $\mathcal{A}$ and that they are independent and identically distributed, with $\mathbb{P}(x_{ij} = a) = p_a$ for some fixed probabilities p_a, $a \in \mathcal{A}$. We assume further that $0, 1 \in \mathcal{A}$ and $p_0, p_1 > 0$. (In the present paper only 0 and 1 have special roles and we might as well assume that $\mathcal{A} = \{0, 1, 2\}$, but because of our application to superboolean rank discussed below, we prefer to state the results for a general $\mathcal{A}$.)

The purpose of the present paper is to study the size of the largest triangular submatrix of X_n, and more precisely its asymptotical behavior as $n \to \infty$. We actually consider four versions of this problem; it turns out that to the first order studied here, they all have the same answer.

Definitions 1.1.

(i) A *submatrix* of a matrix $A = (a_{ij})_{i \in M, j \in N}$ is any matrix obtained by deleting rows and/or columns of A. In other words, it is a matrix $(a_{ij})_{i \in I, j \in J}$ for a non-empty subset of rows $I \subseteq M$ and a non-empty subset of columns $J \subseteq N$. (We preserve the order of the rows and columns in I and J.)

(ii) A *permutation* of a matrix is a matrix obtained by a permutation of the rows and a (possibly different) permutation of the columns. In particular, a *permuted submatrix* of (a_{ij}) is $(a_{i_r j_s})_{r,s=1}^{k,\ell}$ for a sequence of distinct rows $i_1, \dots, i_k$ and a sequence of distinct columns $j_1, \dots, j_\ell$.

(iii) A *(lower) triangular matrix* is a square matrix $(a_{ij})_{i,j=1}^m$ such that $a_{ij} = 0$ for any $i < j$.

Date: 2 September, 2011; revised 13 March, 2013.

2000 *Mathematics Subject Classification.* 03G05, 06E25, 06E75, 60C05.

The research of the first author has been supported by the Israel Science Foundation (ISF grant No. 448/09) and by the Oberwolfach Leibniz Fellows Programme (OWLF), Mathematisches Forschungsinstitut Oberwolfach, Germany.

Acknowledgement: This work was started during a chance meeting of researchers from two different groups at a supper table in Mathematisches Forschungsinstitut Oberwolfach (MFO), Germany in April 2011, and the work was essentially completed during the authors' stay at MFO.

We thank other MFO visitors, in particular Gabor Lugosi, for helpful comments.

We thank the referee for helpful corrections.

(iv) A *special triangular matrix* is a square matrix $(a_{ij})_{i,j=1}^m$ such that $a_{ij} = 0$ for any $i < j$ and $a_{ij} = 1$ when $i = j$. (The remaining entries are arbitrary.)

Note that a $k \times \ell$ submatrix is determined by two *sets* I, J of indices with $|I| = k$, $|J| = \ell$, while a permuted submatrix is determined by two *sequences* $i_1, \dots, i_k$ and $j_1, \dots, j_\ell$ of indices, with each sequence without repetitions.

We define the random variable T_n as the maximal size (= number of rows, or columns) of a submatrix of X_n (an $n \times n$ random matrix as above) that is triangular; similarly T_n^{s}, T_n^{p}, T_n^{ps} denote the maximal sizes of submatrices of X_n that are special triangular, permuted triangular and permuted special triangular, respectively. Note that

$$T_n^{\mathrm{s}} \le T_n \le T_n^{\mathrm{p}} \qquad \text{and} \qquad T_n^{\mathrm{s}} \le T_n^{\mathrm{ps}} \le T_n^{\mathrm{p}}. \tag{1.1}$$

The general motivation for studying these quantities comes from boolean algebra [10] or, more generally, from (tropical) max-plus algebra [1] and supertropical algebra [4]. These algebras take place over semirings and are fundamentally connected to graph theory, in particular square matrices over these semirings correspond uniquely to weighted directed graphs. With this correspondence, basic algebraic notions are naturally substituted by combinatorial ones; for example, the role of the determinant is replaced by the permanent. These combinatorial analogous also help to bypass the lack of negation in the ground structure of semirings. As a consequence, computational complexity, such as computing the rank of a matrix, is not always polynomial and could be NP-complete [11] over this framework.

The specific motivation occurs if one considers either the boolean case ($\mathcal{A} = \{0,1\}$) or the superboolean case ($\mathcal{A} = \{0, 1, 1^\nu\}$) – the simplest example for a supertropical semiring [5, 8]. The latter papers lead to a new algebraic theory of combinatorics, establishing a universal representation of matroids by boolean matrices. In this theory, a square matrix is non-singular if and only if it is permuted special triangular, and the rank of a matrix is thus the maximal size of a permuted special triangular submatrix, see Izhakian and Rhodes [5, 6, 7] for details. Consequently, the rank of the random matrix X_n is T_n^{ps}.

Theorem 1.2. *Let $Q = 1/p_0 > 1$, and let T_n^* be any of $T_n, T_n^{\mathrm{s}}, T_n^{\mathrm{p}}, T_n^{\mathrm{ps}}$. Then, as $n \to \infty$,*

$$T_n^* / \log_Q n \xrightarrow{\mathrm{p}} 2 + \sqrt{2}, \tag{1.2}$$

where $\xrightarrow{\mathrm{p}}$ denotes convergence in probability.

We say that an event occurs *with high probability (w.h.p.)* if its probability tends to 1 as $n \to \infty$. Recall that, by the definition of convergence in probability, (1.2) says that for any $\varepsilon > 0$, w.h.p.

$$(2 + \sqrt{2} - \varepsilon) \log_Q n < T_n^* < (2 + \sqrt{2} + \varepsilon) \log_Q n. \tag{1.3}$$

Furthermore, by (1.1), it suffices to prove the upper inequality for T_n^{p} and the lower for T_n^{s}. The upper inequality is proved in Section 2 and the lower in Section 3; the proofs are based on the first and second moment methods. (See e.g. [9, p. 54] for a general description of these methods.)

Remark 1.3. The corresponding problem of the largest square submatrix with only 0's (or, equivalently, after interchange of 0 and 1, with only 1's) has been studied by several authors, see [13] and the references therein. It is shown in [13] that if S_n is the size of the largest such matrix, then $S_n/\log_Q n \xrightarrow{\mathrm{p}} 2$. This problem can be seen as finding the largest balanced complete subgraph of a random bipartite graph. The analogous problem of finding the largest complete set in a random graph $G(n,p)$ (or, equivalently, the largest independent set in $G(n,1-p)$) was solved by Bollobás and Erdős [3] and Matula [12], see also [2] and [9]; again the size, C_n say, is asymptotically $2\log_Q n$, where $Q = 1/p$. (We have no intuitive explanation for the extra summand $\sqrt{2}$ in the triangular case. Note also that the number of 0's in the largest triangular submatrix is $\approx \frac{1}{2}(2+\sqrt{2})^2\log_Q^2 n = (3+2\sqrt{2})\log_Q^2 n$, which is larger than for the largest square submatrix with only zeros where the number is $\approx 4\log_Q^2 n$.)

Note that $T_n \geq S_n \geq \lfloor T_n^{\mathrm{p}}/2\rfloor \geq \lfloor T_n/2\rfloor$, which shows that T_n, T_n^{p} and S_n are equal within a factor of $2+o(1)$, and in particular of the same order of magnitude. However, it does not seem possible to get the right constant in front of $\log_Q n$ for one of these problems from the other.

For the largest square zero submatrix and the largest cliques in $G(n,p)$, much more precise estimates are known, see [13] and [2, 9]; for example, it follows that if

$$s(n) = 2\log_Q n - 2\log_Q \log_Q n + 2\log_Q(e/2),$$

then for any $\varepsilon > 0$, $\lfloor s(n)-\varepsilon\rfloor \leq S_n \leq \lfloor s(n)+\varepsilon\rfloor$ and $\lfloor s(n)+1-\varepsilon\rfloor \leq C_n \leq \lfloor s(n)+1+\varepsilon\rfloor$ w.h.p. (and, in fact, almost surely); in particular the sizes are concentrated on one or at most two values. It would be interesting to find similar sharper versions of the result above, which leads to the following open problems.

Problem 1.4. Find second order terms for $T_n, T_n^{\mathrm{s}}, T_n^{\mathrm{p}}, T_n^{\mathrm{ps}}$, and if possible even sharper estimates, and see if they differ between the four versions. In particular, what are the orders of the differences $T_n^{\mathrm{p}} - T_n$, $T_n - T_n^{\mathrm{s}}$, ...?

Problem 1.5. Are the quantities $T_n, T_n^{\mathrm{s}}, T_n^{\mathrm{p}}, T_n^{\mathrm{ps}}$ concentrated on at most two values each?

Problem 1.6. Prove a version of Theorem 1.2 (or a stronger result) with convergence almost surely instead of just in probability, seeing X_n as submatrices of an infinite random matrix in the natural way.

Problem 1.7. Find corresponding results when p_0 and p_1 depend on n. The case when p_0 tends to 1 (not too fast) seems to be the most interesting.

Remark 1.8. We consider for simplicity only square matrices X_n, but the definitions extend to general $m \times n$ matrices. Since the quantities $T_n, T_n^{\mathrm{s}}, T_n^{\mathrm{p}}, T_n^{\mathrm{ps}}$ are monotone if we add rows or columns, the result of Theorem 1.2 holds as long as $\log m/\log n \to 1$; this includes for example the case $m/n \to c \in (0,\infty)$. We have not investigated other cases such as $m = n^\gamma$ for some $\gamma > 0$.

1.1. **Notation.** We let $\lfloor x\rfloor$ and $\lceil x\rceil$ denote the largest and smallest integers such that $\lfloor x\rfloor \leq x \leq \lceil x\rceil$. We write $[m,n]$ for the interval $\{m, m+1, \dots, n\}$ of integers between m and n. Further, log denotes the natural logaritm $\log_e$; recall that $\log_Q n = \log n/\log Q$. Henceforth, an $n\times n$ random matrix $X_n = (x_{ij})_{i,j=1}^n$ is denoted as X, for short.

2. Proof of upper bound

As said above, it suffices to show that $T_n^{\mathrm{p}} \leq (2+\sqrt{2}+\varepsilon)\log_Q n$ w.h.p. for every $\varepsilon > 0$. We will use the first moment method, i.e., show that a suitable expectation tends to 0. However, for reasons discussed below, we will not obtain the right constant by calculating the expected number of (permuted) triangular submatrices of X. Instead we consider the following type of submatrices.

Definition 2.1. Let $1 \le \ell \le k$. A (k,ℓ)-corner matrix is an $\ell \times \ell$ matrix $(a_{ij})_{i,j=1}^{\ell}$ such that

$$a_{ij} = 0 \qquad \text{if} \qquad i < j + k - \ell; \tag{2.1}$$

if further $a_{ij} = 1$ when $i = j + k - \ell$, the matrix is a *special* (k,ℓ)*-corner matrix.*

Thus the $\ell \times \ell$ submatrix in the upper right corner of a [special] lower $k \times k$ triangular matrix is a [special] (k,ℓ)-corner matrix, and conversely. Note that if $\ell \le k/2$, then a (k,ℓ)-corner matrix is 0, and if $\ell = k$ then a (k,ℓ)-corner matrix is the same as a triangular matrix.

Let $\nu_0(k,\ell)$ be the number of entries required to be 0 by (2.1). Thus $\nu_0(k,\ell) = \ell^2$ when $\ell \le k/2$; for $\ell \ge k/2$ we have

$$\begin{aligned}
\nu_0(k,\ell) &= \sum_{j=1}^{\ell} \min(j + k - \ell - 1, \ell) \\
&= \sum_{j=1}^{2\ell-k} (j + k - \ell - 1) + \sum_{j=2\ell-k+1}^{\ell} \ell \\
&= \frac{(2\ell-k)(2\ell-k+1)}{2} + (2\ell-k)(k-\ell-1) + (k-\ell)\ell \\
&= \frac{4k\ell - k^2 - 2\ell^2 + k - 2\ell}{2}.
\end{aligned} \tag{2.2}$$

Similarly, let $\nu_1(k,\ell)$ be the number of entries required to be 1 in a special (k,ℓ)-corner matrix. Thus $\nu_1(k,\ell) = 0$ when $\ell \le k/2$ and $\nu_1(k,\ell) = 2\ell - k$ when $\ell \ge k/2$. Further, let $\nu(k,\ell) = \nu_0(k,\ell) + \nu_1(k,\ell)$ be the total number of fixed entries in a special (k,ℓ)-corner matrix. If $\ell \ge k/2$, then by (2.2)

$$\nu(k,\ell) = \frac{4k\ell - k^2 - 2\ell^2 - k + 2\ell}{2}. \tag{2.3}$$

Let $1 \le \ell \le m$ and let $Y_{m,\ell}$ be the number of permuted (m,ℓ)-corner submatrices in X. Note that if X contains a permuted triangular $m \times m$ submatrix A, then a suitable submatrix of A is a permuted (m,ℓ)-corner submatrix of X. Hence, if $T_n^{\mathrm{p}} \ge m$, then $Y_{m.\ell} \ge 1$, and Markov's inequality yields

$$\mathbb{P}(T_n^{\mathrm{p}} \ge m) \le \mathbb{P}(Y_{m,\ell} \ge 1) \le \mathbb{E}\, Y_{m,\ell}\ . \tag{2.4}$$

The expected value $\mathbb{E}\, Y_{m,\ell}$ is easily computed. The number of permuted $\ell \times \ell$ submatrices of X is $(n)_\ell \cdot (n)_\ell$, where $(n)_\ell = n(n-1)\cdots(n-\ell+1)$, and for each such matrix, the probability that it is an (m,ℓ)-corner matrix is $p_0^{\nu_0(m,\ell)}$, with $\nu_0(m,\ell)$ given above. Thus,

$$\mathbb{E}\, Y_{m,\ell} = (n)_\ell^2 \cdot p_0^{\nu_0(m,\ell)} \le \exp\bigl(2\ell \log n - \log Q \cdot \nu_0(m,\ell)\bigr). \tag{2.5}$$

Taking $m = \lceil s \log n \rceil$ and $\ell = \lceil t \log n \rceil$ for some fixed s and t with $s/2 < t \le s$, we have by (2.5) and (2.2),

$$\mathbb{E}\, Y_{m,\ell} \le \exp\bigl(2t(\log n)^2 - \log Q \cdot (2st - s^2/2 - t^2)(\log n)^2 + O(\log n)\bigr). \tag{2.6}$$

We see from (2.6) that if we choose s and t such that $s/2 < t \le s$ and

$$2t - \log Q \cdot (2st - s^2/2 - t^2) < 0, \tag{2.7}$$

then $\mathbb{E}\, Y_{m,\ell} \to 0$ and thus by (2.4)

$$\mathbb{P}(T_n^{\mathrm{p}} \geq s \log n) = \mathbb{P}(T_n^{\mathrm{p}} \geq m) \leq \mathbb{E}\, Y_{m,\ell} \to 0; \tag{2.8}$$

hence $T_n^{\mathrm{p}} < s \log n$ w.h.p.

Write for convenience $\gamma = 1/\log Q$. The left hand side of (2.7) is, for a fixed s, maximized when $t = s - \gamma$, and then its value is, by a short calculation,

$$2s - \gamma - \frac{s^2}{2\gamma} = -\frac{s^2 - 4s\gamma + 2\gamma^2}{2\gamma} = -\frac{(s-2\gamma)^2 - 2\gamma^2}{2\gamma},$$

which is negative for $s > 2\gamma + \sqrt{2}\gamma$. Consequently, taking any $s > (2 + \sqrt{2})\gamma$ and $t = s - \gamma$, which clearly satisfies $s/2 < t < s$, (2.8) yields $T_n^{\mathrm{p}} < s \log n$ w.h.p. It remains only to note that $\gamma \log n = \log n / \log Q = \log_Q n$.

Remark 2.2. If we instead estimate the number of (permuted) triangular submatrices, we are taking $\ell = m$ and $t = s$ in the calculations above and we only obtain the weaker estimate $T_n^{\mathrm{p}} \leq (4+\varepsilon) \log_Q n$ w.h.p. The reason that the first moment method does not yield a sharp estimate in this case is that triangular submatrices of large size tend to occur in large clusters; thus the expected number of such submatrices of a given size can tend to infinity although the probability that the number is nonzero tends to 0. See also the proof of the lower bound in Section 3, which shows that a (k,ℓ)-corner matrix of close to maximal size w.h.p. can be extended to a triangular submatrix in many different ways.

3. Proof of lower bound

We begin by stating three lemmas; the first is elementary and the two others contain the main probabilistic arguments. The proofs are provided later.

Lemma 3.1. *Suppose that $k_1 \geq \ell_1 \geq 1$, $k_2 \geq \ell_2 \geq 1$, and $2(\ell_1 - \ell_2) \geq k_1 - k_2 \geq 0$. Then every special (k_1, ℓ_1)-corner matrix contains a special (k_2, ℓ_2)-corner submatrix.*

Lemma 3.2. *Let $\varepsilon > 0$. There exists some $k = k(n)$ and $\ell = \ell(n)$ with*

$$(2 + \sqrt{2} - \varepsilon) \log_Q n \leq k \leq (2 + \sqrt{2}) \log_Q n,$$

and

$$(1 + \sqrt{2} - \varepsilon) \log_Q n \leq \ell \leq (1 + \sqrt{2}) \log_Q n$$

such that w.h.p. X contains a special (k,ℓ)-corner submatrix.

Lemma 3.3. *Let X' be the submatrix $(x_{ij})_{i \leq n/2,\, j > n/2}$ comprising the upper right quarter of X. Let $\varepsilon > 0$ and let $k = k(n)$ and $\ell = \ell(n)$ be such that $k/2 < \ell < k$ and $k - \ell \leq (1-\varepsilon) \log_Q n$. If X' contains a special (k,ℓ)-corner submatrix, then w.h.p. X contains a special triangular $k \times k$ submatrix, and thus $T_n^{\mathrm{s}} \geq k$.*

Proof of lower bound in Theorem 1.2. Let $0 < \varepsilon < 1/3$. Let X' be the upper right quarter of X as in Lemma 3.3. By Lemma 3.2, there exists k_1 and ℓ_1 with

$$(2 + \sqrt{2} - \varepsilon) \log_Q \lfloor n/2 \rfloor \leq k_1 \leq (2 + \sqrt{2}) \log_Q \lfloor n/2 \rfloor,$$
$$(1 + \sqrt{2} - \varepsilon) \log_Q \lfloor n/2 \rfloor \leq \ell_1 \leq (1 + \sqrt{2}) \log_Q \lfloor n/2 \rfloor$$

such that there w.h.p. is a special (k_1, ℓ_1)-corner submatrix M_1 of X'.

Note that $k_1 - \ell_1 \leq (1+\varepsilon) \log_Q n$. Let $d = \lceil 2\varepsilon \log_Q n \rceil$, $k = k_1 - 2d$, and $\ell = \ell_1 - d$. By Lemma 3.1, there is a special (k,ℓ)-corner submatrix M_2 of M_1. It is easily verified that k and ℓ satisfy the conditions of Lemma 3.3, and thus Lemma 3.3 shows that w.h.p.

$$T_n^{\mathrm{s}} \geq k \geq (2+\sqrt{2}-5\varepsilon)\log_Q n + O(1).$$

The bound $T_n^{\mathrm{s}} \geq (2+\sqrt{2}-\varepsilon)\log_Q n$ w.h.p. follows by replacing ε by $\varepsilon/6$. This completes the proof of Theorem 1.2 since $T_n^* \geq T_n^{\mathrm{s}}$ by (1.1). □

It remains to prove the lemmas.

Proof of Lemma 3.1. Let A be a special (k,ℓ)-corner matrix. The submatrix obtained by deleting the first row and last column is a special $(k-2,\ell-1)$-corner matrix. Similarly, we obtain a special $(k-1,\ell-1)$-corner matrix by deleting the last row and last column, and a special $(k,\ell-1)$-corner matrix by deleting the last row and first column.

The lemma now follows by induction on $\ell_1-\ell_2$. □

Proof of Lemma 3.2. We may assume that $\varepsilon < 1/4$. We consider a block version of (k,ℓ)-corner matrices.

Let N be a large integer and let $K = \lceil (2+\sqrt{2}-\varepsilon)N\rceil$ and $L = \lceil (1+\sqrt{2}-\varepsilon)N\rceil = K-N$; note that $K > L > K/2$. Let $n_1 = \lfloor n/L \rfloor$ and divide the interval $[1,n]$ into the L subintervals $E_i = [(i-1)n_1+1, in_1]$, $i=1,\dots,L$, ignoring the possible remainder at the end. Let X_{ij} be the $n_1\times n_1$ submatrix $(x_{rs})_{r\in E_i,\, s\in E_j}$ of X.

Let

$$q = \lceil N^{-1}\log_Q n\rceil \tag{3.1}$$

and consider the submatrices of X obtained by choosing q rows from each E_i and q columns from each E_j, $i,j=1,\dots,L$. We denote the set of all such submatrices by $\mathcal{M}$; each $M\in\mathcal{M}$ is identified by its set of rows and columns, and the number of them is thus

$$|\mathcal{M}| = \binom{n_1}{q}^{2L}. \tag{3.2}$$

Each M is a $Lq\times Lq$ submatrix of X which consists of L^2 blocks M_{ij}, $i,j\in\{1,\dots,L\}$, where M_{ij} is a $q\times q$ submatrix of X_{ij}.

We say that the submatrix $M\in\mathcal{M}$ is *good* (for a given realization of the random matrix X) if $M_{ij}=0$ when $i<j+K-L$ and $M_{ij}=I$ (the $q\times q$ identity matrix) when $i=j+K-L$; otherwise M is called *bad*. Thus, a good submatrix can be seen as a special (K,L)-corner matrix of $q\times q$ matrices.

Note that a good submatrix M is a special (Kq,Lq)-corner matrix, and that $k=Kq$ and $\ell=Lq$ satisfy the inequalities in the lemma if N and q are large enough. Hence it suffices to show that if N is large enough, then there exists w.h.p. at least one good submatrix $M\in\mathcal{M}$.

Let I_M be the indicator that M is good, i.e., $I_M=1$ if M is good and $I_M=0$ if M is bad, and let $Z=\sum_{M\in\mathcal{M}} I_M$ be the number of good submatrices $M\in\mathcal{M}$. Our task is to show that $Z\geq 1$ w.h.p., which we do by estimating the mean and variance.

In order for M to be good, the number of submatrices M_{ij} required to be 0 is $\nu_0(K,L)$, and the number required to be I is $\nu_1(K,L)$. Consequently, the number of entries required to be 0 is $\nu_0(K,L)q^2+\nu_1(K,L)(q^2-q)=\nu(K,L)q^2-\nu_1(K,L)q$ and the number of entries required to be 1 is $\nu_1(K,L)q$. Hence, denoting the probability that M is good by π, for each $M\in\mathcal{M}$,

$$\pi = \mathbb{P}(I_M=1) = p_0^{\nu(K,L)q^2-\nu_1(K,L)q}\,p_1^{\nu_1(K,L)q}. \tag{3.3}$$

We have by (2.3), recalling $K = L + N$,

$$\begin{aligned}\nu(K,L) &= \frac{4(L+N)L - (L+N)^2 - 2L^2 + O(N)}{2}\\ &= \frac{L^2 + 2LN - N^2}{2} + O(N)\\ &= \Bigl(2 + 2\sqrt{2} - (2+\sqrt{2})\varepsilon + \frac{\varepsilon^2}{2}\Bigr)N^2 + O(N)\\ &< (2 - \varepsilon/2)LN, \end{aligned} \tag{3.4}$$

provided N is chosen large enough. We fix such an N; thus K and L are now fixed, while $n \to \infty$. By (3.1),

$$\log n = \log_Q n \cdot \log Q = Nq \log Q + O(1). \tag{3.5}$$

Furthermore, (3.1) also yields, as $n \to \infty$, $q \le \log_Q n \ll n_1$. Hence, by Stirling's formula,

$$\log \binom{n_1}{q} = q \log n_1 + O\Bigl(\frac{q^2}{n_1}\Bigr) - \log(q!) = q \log n + O(q \log q).$$

Consequently, by (3.2), (3.3), (3.4) and (3.5),

$$\begin{aligned}\mathbb{E}\, Z &= |\mathcal{M}|\, \mathbb{P}(I_M = 1) = |\mathcal{M}|\pi\\ &= \exp\Bigl(2L\bigl(q \log n + O(q \log q)\bigr) - \nu(K,L) q^2 \log Q + O(q)\Bigr)\\ &\ge \exp\Bigl(2L\bigl(q \log n\bigr) - (2 - \varepsilon/2)LNq^2 \log Q + O(q \log q)\Bigr)\\ &= \exp\Bigl((\varepsilon LN \log Q/2)q^2 + O(q \log q)\Bigr) \to \infty. \end{aligned} \tag{3.6}$$

To estimate the variance $\operatorname{Var}(Z)$, we first calculate the covariance $\operatorname{Cov}(I_M, I_{M'}) = \mathbb{E}(I_M I_{M'}) - \mathbb{E}(I_M)\,\mathbb{E}(I_{M'})$ for two submatrices $M, M' \in \mathcal{M}$. Let a_i be the number of common rows in E_i of M and M', and let b_j be the number of common columns in E_j. Then M_{ij} has $a_i b_j$ entries in common with M'_{ij}, so their union has $2q^2 - a_i b_j$ elements.

For $i < j + K - L$, we have

$$\frac{\mathbb{P}(M_{ij} = 0 = M'_{ij})}{\mathbb{P}(M_{ij} = 0)\,\mathbb{P}(M'_{ij} = 0)} = \frac{p_0^{2q^2 - a_i b_j}}{p_0^{2q^2}} = p_0^{-a_i b_j}. \tag{3.7}$$

For $i = j + K - L$, we want $M_{ij} = M'_{ij} = I$, so we have to consider also the required positions of the 1's in M_{ij} and M'_{ij}. In many cases, the rows and columns chosen for M_{ij} and M'_{ij} are such that the conditions $M_{ij} = I$ and $M'_{ij} = I$ are contradictory, so $\mathbb{P}(M_{ij} = M'_{ij} = I) = 0$. Otherwise, the $a_i b_j$ common entries of M_{ij} and M'_{ij} contain some number of entries, d say, that have to be 1 in both M_{ij} and M'_{ij}, while the remaining $a_i b_j - d$ have to be 0 in both, and then

$$\frac{\mathbb{P}(M_{ij} = M'_{ij} = I)}{\mathbb{P}(M_{ij} = I)\,\mathbb{P}(M'_{ij} = I)} = p_0^{-(a_i b_j - d)} p_1^{-d} = p_0^{-a_i b_j} \Bigl(\frac{p_0}{p_1}\Bigr)^d; \tag{3.8}$$

note that $0 \le d \le \min(a_i, b_j)$. Combining (3.7) and (3.8) by taking the product over all pairs (i,j) with $i \le j + K - L$, and recalling that $K - L = N$, we obtain the upper bound

$$\frac{\mathbb{P}(I_M = I_{M'} = 1)}{\mathbb{P}(I_M = 1)\,\mathbb{P}(I_{M'} = 1)} \le p_0^{-\sum_{i,j: i \le j+N} a_i b_j} \max\Bigl(\Bigl(\frac{p_0}{p_1}\Bigr)^{L \sum_i a_i}, 1\Bigr). \tag{3.9}$$

Let $\pi = \mathbb{P}(I_M = 1)$, $C_1 = \max\{(p_0/p_1)^L, 1\}$ and, for a given pair M, M', $A = \sum_i a_i$ and $B = \sum_j b_j$, be the numbers of common rows and columns, respectively, of M and M'. Then (3.9) yields

$$\operatorname{Cov}(I_M, I_{M'}) \le \mathbb{P}(I_M = I_{M'} = 1) \le Q^{\sum_{i,j:i\le j+N} a_i b_j} C_1^A \pi^2. \tag{3.10}$$

Let

$$\tau = \tau\big((a_i),(b_j)\big) = \sum_{i,j:i\le j+N} a_i b_j \tag{3.11}$$

and let $\tau(A,B)$ be the maximum of τ for given sums $A = \sum_i a_i$ and $B = \sum_j b_j$, with $a_i, b_j \in [0,q]$. If $i_1 < i_2$ and we increase a_{i_1} by some Δ to $a_{i_1} + \Delta$ and decrease a_{i_2} by the same Δ to $a_{i_2} - \Delta$, then $\tau = \sum_{i,j:i\le j+N} a_i b_j$ cannot decrease. The same happens if we decrease b_{j_1} and increase b_{j_2} with $j_1 < j_2$. Consequently, given A and B, the sum τ is maximized when, for some indices $i_*, j_* \in [1, L]$,

$$a_i = q \text{ when } i < i_*, \qquad a_i = 0 \text{ when } i > i_*; \tag{3.12}$$
$$b_j = 0 \text{ when } j < j_*, \qquad b_j = q \text{ when } j > j_*. \tag{3.13}$$

Returning to (3.10), we have the estimate $\operatorname{Cov}(I_M, I_{M'}) \le Q^{\tau(A,B)} C_1^A \pi^2$. If $A = 0$ or if $B = 0$, then M and M' are disjoint submatrices of X, and thus independent, so in this case $\operatorname{Cov}(I_M, I_{M'}) = 0$. Consequently,

$$\operatorname{Var}(Z) = \sum_{M,M'} \operatorname{Cov}(I_M, I_{M'}) \le \sum_{M,M':A,B>0} Q^{\tau(A,B)} C_1^A \pi^2, \tag{3.14}$$

where A and B are defined as above, given M and M'.

For a given $M \in \mathcal{M}$, the number of submatrices $M' \in \mathcal{M}$ with given $a_1, \dots, a_L$, $b_1, \dots, b_L$ is

$$N\big((a_i)_i, (b_j); q\big) = \prod_{i=1}^L \binom{q}{a_i}\binom{n_1 - q}{q - a_i} \prod_{j=1}^L \binom{q}{b_j}\binom{n_1 - q}{q - b_j}.$$

We have, for any $a \in [0, q]$,

$$\frac{\binom{q}{a}\binom{n_1-q}{q-a}}{\binom{n_1}{q}} \le \frac{q^a\binom{n_1-a}{q-a}}{\binom{n_1}{q}} = q^a \prod_{i=0}^{a-1} \frac{q-i}{n_1-i} \le q^a\Big(\frac{q}{n_1}\Big)^a = \Big(\frac{q^2}{n_1}\Big)^a. \tag{3.15}$$

Thus, recalling (3.2),

$$\frac{N\big((a_i)_i, (b_j); q\big)}{|\mathcal{M}|} \le \Big(\frac{q^2}{n_1}\Big)^{A+B}.$$

Moreover, given A and B, the number of choices of $a_1, \dots, a_L$ with sum A is $\le (A+1)^L \le 2^{AL}$, and similarly the number of $b_1, \dots, b_L$ is $\le 2^{BL}$. Hence, for each $M \in \mathcal{M}$, the number of M' with given A and B is at most, using (3.15),

$$2^{AL} 2^{BL} \Big(\frac{q^2}{n_1}\Big)^{A+B} |\mathcal{M}| = \Big(\frac{2^L q^2}{n_1}\Big)^{A+B} |\mathcal{M}| \le \Big(\frac{C_2 q^2}{n}\Big)^{A+B} |\mathcal{M}|,$$

where $C_2 = (L+1)2^L$ (for n large enough). Since M can be chosen in $|\mathcal{M}|$ ways, and $A, B \le Lq$, (3.14) yields, recalling $\mathbb{E}\, Z = |\mathcal{M}|\pi$,

$$\operatorname{Var}(Z) \le \sum_{A,B=1}^{Lq} |\mathcal{M}| \Big(\frac{C_2 q^2}{n}\Big)^{A+B} |\mathcal{M}| Q^{\tau(A,B)} C_1^{A+B} \pi^2$$
$$= (\mathbb{E}\, Z)^2 \sum_{A,B=1}^{Lq} \Big(\frac{C_3 q^2}{n}\Big)^{A+B} Q^{\tau(A,B)}, \tag{3.16}$$

with $C_3 = C_1 C_2$. We write (3.16) as $\operatorname{Var}(Z) = (\mathbb{E}\, Z)^2 \sum_{A,B} \lambda(A,B)$, with

$$\lambda(A,B) = \Big(\frac{C_3 q^2}{n}\Big)^{A+B} Q^{\tau(A,B)}. \tag{3.17}$$

Claim. *If $A, B \in [1, Lq]$, then $\lambda(A,B) \le \max\{\lambda(1,1), \lambda(Lq,Lq)\}$; in other words, $\lambda(A,B)$ attains its maximum for $A = B = 1$ or $A = B = Lq$.*

To prove the claim, let (a_i) and (b_j) be vectors that maximize τ in (3.11) for some given A and B; we may thus assume that (3.12) and (3.13) hold. We first note that if $A < Nq$, then by (3.12) we have $i_* \le N$ and $a_i = 0$ when $i > N$; hence

$$\tau(A,B) = \tau = \sum_{i,j:i\le j+N} a_i b_j = \sum_{i,j=1}^{L} a_i b_j = AB$$

and thus

$$\lambda(A,B) = (C_3 q^2/n)^{A+B} Q^{AB}.$$

Keeping A fixed, this is maximized by either $B = 1$ or $B = Lq$.

On the other hand, if $A \ge Nq$, then (3.12) yields $a_i = q$ when $i \le N$. Hence, increasing any b_j by 1 will increase τ in (3.11) by $\sum_{i:i\le j+N} a_i \ge Nq$, and thus $\tau(A, B+1) \ge \tau(A,B) + Nq$. Consequently, by (3.17) and (3.1),

$$\frac{\lambda(A,B+1)}{\lambda(A,B)} = \Big(\frac{C_3 q^2}{n}\Big) Q^{\tau(A,B+1)-\tau(A,B)} \ge \Big(\frac{C_3 q^2}{n}\Big) Q^{Nq} \ge C_3 q^2 > 1,$$

and thus $\lambda(A,B) \le \lambda(A, Lq)$ for any $B \le Lq$.

Hence, for any fixed $A \le Lq$, $\lambda(A,B)$ is maximized by either $B = 1$ or $B = Lq$. By symmetry, for fixed B, the maximum is attained for $A = 1$ or $A = Lq$. Consequently, the maximum for all $A, B \in [1, Lq]$ is attained for $A, B \in \{1, Lq\}$. Moreover, $\lambda(1, Lq) = \lambda(Lq, 1)$ by symmetry and $\lambda(Lq,1) \le \lambda(Lq,Lq)$ by the case $A \ge Nq$ above. Hence, the claim follows.

Having proved the claim, we calculate easily the two extreme cases in it. For $A = B = 1$, $\tau(1,1) = 1$ and

$$\lambda(1,1) = \Big(\frac{C_3 q^2}{n}\Big)^2 Q = O\Big(\frac{\log^4 n}{n^2}\Big). \tag{3.18}$$

For $A = B = Lq$, all $a_i = b_j = q$, and thus $\tau(Lq, Lq) = \nu(K, L)q^2$, with $\nu(K, L)$ given by (2.3). Hence, recalling $q = O(\log n)$, (3.5) and (3.4),

$$\begin{aligned}
\lambda(Lq, Lq) &= \Bigl(\frac{C_3 q^2}{n}\Bigr)^{2Lq} Q^{\nu(K,L)q^2} \\
&= \exp\Bigl(-2Lq \log n + O(q \log q) + \nu(K, L)q^2 \log Q\Bigr) \\
&= \exp\Bigl((-2LNq^2 + \nu(K, L)q^2) \log Q + O(q \log q)\Bigr) \\
&\le \exp\Bigl(-(\varepsilon LN \log Q/2)q^2 + O(q \log q)\Bigr). \qquad (3.19)
\end{aligned}$$

For large n, this is less than $\exp(-2Nq) < n^{-2}$. Consequently, the claim and (3.18)–(3.19) shows that for all $A, B \le Lq$,

$$\lambda(A, B) = O\Bigl(\frac{\log^4 n}{n^2}\Bigr). \qquad (3.20)$$

Finally, by (3.16) and (3.20),

$$\frac{\operatorname{Var}(Z)}{(\mathbb{E}\, Z)^2} \le \sum_{A,B=1}^{Lq} \lambda(A, B) = O\Bigl(\frac{q^2 \log^4 n}{n^2}\Bigr) = O\Bigl(\frac{\log^6 n}{n^2}\Bigr) = o(1), \qquad (3.21)$$

as $n \to \infty$. This is what we need: by Chebyshev's inequality

$$\mathbb{P}(Z = 0) \le \frac{\operatorname{Var}(Z)}{(\mathbb{E}\, Z)^2};$$

hence (3.21) yields $\mathbb{P}(Z = 0) \to 0$, and thus $Z \ge 1$ w.h.p., which completes the proof. $\square$

Proof of Lemma 3.3. Condition on X' and fix a special (k, ℓ)-corner submatrix $M' = (x_{i'_r,j'_s})_{r,s=1}^{\ell}$ of X'; thus $1 \le i'_1 < \cdots < i'_\ell \le n/2$ and $n/2 < j'_1 < \cdots < j'_\ell \le n$. We try to complete M' to a $k \times k$ special triangular matrix by adding $k - \ell$ columns $j_1 < \cdots < j_{k-\ell} \le n/2$ in the left half and $k - \ell$ rows $n/2 < i_1 < \cdots < i_{k-\ell} \le n$ in the lower half of X; we do this by trying the columns one by one until we find first a suitable j_1 (i.e., one with $x_{i'_1 j_1} = 1$), then a suitable j_2 (one with $x_{i'_1 j_2} = 0$ and $x_{i'_2 j_2} = 1$), and so on until $j_{k-\ell}$, and similarly for $i_1, \dots, i_{k-\ell}$. (Note that we only search among the rows and columns that do not intersect X'.)

Let $r \le k - \ell$. Each time we try a column in order to find j_r, we want one specific entry in it to be 1 and $r - 1$ others to be 0; the probability of this is $\pi_r = p_0^{r-1} p_1$, independently of X' and what has happened earlier. If T_r is the number of columns that we have to try until we find j_r, then T_r thus has a geometric distribution

$$\mathbb{P}(T_r = t) = (1 - \pi_r)^{t-1} \pi_r, \qquad t = 1, 2, \dots.$$

This distribution has mean $\mathbb{E}\, T_r = 1/\pi_r$ and variance $\operatorname{Var} T_r = (1 - \pi_r)/\pi_r^2$; hence the sum $S := T_1 + \cdots + T_{k-\ell}$ has mean

$$\mathbb{E}\, S = \sum_{r=1}^{k-\ell} \mathbb{E}\, T_r = \sum_{r=1}^{k-\ell} \pi_r^{-1} = \sum_{r=1}^{k-\ell} p_1^{-1} Q^{r-1} = O\bigl(Q^{k-\ell}\bigr) = O\bigl(n^{1-\varepsilon}\bigr) = o(n)$$

and variance

$$\operatorname{Var} S = \sum_{r=1}^{k-\ell} \operatorname{Var} T_r \le \sum_{r=1}^{k-\ell} \pi_r^{-2} = O\big(Q^{2(k-\ell)}\big) = O\big(n^{2(1-\varepsilon)}\big) = o(n^2).$$

The search for $j_1, \dots, j_{k-\ell}$ succeeds if $S \le n/2$. Consequently the probability of failure is, using Chebyshev's inequality, for n so large that $\mathbb{E}\, S < n/4$,

$$\mathbb{P}(S > n/2) \le \frac{\operatorname{Var} S}{(n/2 - \mathbb{E}\, S)^2} \le \frac{\operatorname{Var} S}{(n/4)^2} = o(1).$$

Hence, w.h.p. we succeed and find suitable columns $j_1, \dots, j_{k-\ell}$; similarly w.h.p. we find also suitable rows $i_1, \dots, i_{k-\ell}$, and we can extend M' to a special triangular $k \times k$ matrix. □

Note that w.h.p. S is much less than $n/2$, so we have a wide margin in this proof and there are w.h.p. many different choices of rows and columns that work, and thus many different ways to extend M' to a special triangular matrix, cf. Remark 2.2.

References

[1] M. Akian, R. Bapat, and S. Gaubert, Max-plus algebra. In: Hogben, L., Brualdi, R., Greenbaum, A., Mathias, R. (eds.) *Handbook of Linear Algebra*. Chapman and Hall, London, 2006.

[2] B. Bollobás, *Random Graphs*, 2nd ed. Cambridge Univ. Press, Cambridge, 2001.

[3] B. Bollobás and P. Erdős, Cliques in random graphs. *Math. Proc. Camb. Phil. Soc.* **80** (1976), 419–427.

[4] Z. Izhakian, Tropical arithmetic and tropical matrix algebra. *Commun. in Alg.*, **37**(4) (2009), 1445–1468.

[5] Z. Izhakian and J. Rhodes, New representations of matroids and generalizations. Preprint, 2011. `arXiv:1103.0503`.

[6] Z. Izhakian and J. Rhodes, Boolean representations of matroids and lattices. Preprint, 2011. `arXiv:1108.1473`.

[7] Z. Izhakian and J. Rhodes, C-dependence and c-rank of posets and lattices. Preprint, 2011. `arXiv:1110.3553`.

[8] Z. Izhakian and L. Rowen, Supertropical algebra. *Adv. in Math*, **225**(8) (2010), 2222–2286.

[9] S. Janson, T. Łuczak & A. Ruciński, *Random Graphs*. Wiley, New York, 2000.

[10] K. H. Kim, *Boolean Matrix Theory and Applications*. Marcel Dekker, New York, 1982.

[11] K. H. Kim and F. W. Roush, Kapranov rank vs. tropical rank. *Proc. Amer. Math. Soc*, **134** (2006), 2487–2494.

[12] D. Matula, The largest clique size in a random graph. Tech. Rep., Dept. Comp. Sci., Southern Methodist Univerity, Dallas, Texas, 1976.

[13] X. Sun and A. B. Nobel, On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res.* **9** (2008), 2431–2453.

School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

Department of Mathematics, Bar-Ilan University, Ramat-Gan 52900, Israel.
E-mail address: zzur@math.biu.ac.il

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
E-mail address: svante.janson@math.uu.se
URL: http://www2.math.uu.se/∼svante/

Department of Mathematics, University of California, Berkeley, 970 Evans Hall #3840, Berkeley, CA 94720-3840 USA.
E-mail address: blvdbastille@aol.com;rhodes@math.berkeley.edu