The Minrank of Random Graphs

Alexander Golovnev, *Yahoo Research,*
Oded Regev, *Courant Institute of Mathematical Sciences, New York University,*
and Omri Weinstein, *Columbia University*

***Abstract*—The *minrank* of a directed graph G is the minimum rank of a matrix M that can be obtained from the adjacency matrix of G by switching some ones to zeros (i.e., deleting edges) and then setting all diagonal entries to one. This quantity is closely related to the fundamental information-theoretic problems of (linear) *index coding* (Bar-Yossef et al., IEEE Trans. Inf. Theory 2011), network coding (Effros et al., IEEE Trans. Inf. Theory 2015) and distributed storage (Mazumdar, ISIT, 2014).**

We prove tight bounds on the minrank of directed Erdős-Rényi random graphs $G(n,p)$ for all regimes of $p \in [0,1]$. In particular, for any constant p, we show that $\mathsf{minrk}(G) = \Theta(n/\log n)$ with high probability, where G is chosen from $G(n,p)$. This bound gives a near quadratic improvement over the previous best lower bound of $\Omega(\sqrt{n})$ (Haviv and Langberg, ISIT 2012), and partially settles an open problem raised by Lubetzky and Stav (IEEE Trans. Inf. Theory 2009). Our lower bound matches the well-known upper bound obtained by the "clique covering" solution, and settles the linear index coding problem for random knowledge graphs.

***Index Terms*— Index coding, Minrank, Linear index coding.**

I. Introduction

In the *index coding* problem ([2], [3]), a sender wishes to *broadcast*, over a noiseless channel, an n-symbol string $x \in \mathbb{F}^n$ (where $\mathbb{F}$ is a finite field) to a group of n receivers $R_1, \ldots, R_n$, each equipped with some *side information*, namely, a subvector x_{K_i} of x (indexed by a subset $K_i \subseteq \{1, \ldots, n\}$). The index coding problem asks what is the minimum length m of a broadcast message that allows each receiver R_i to retrieve the ith symbol x_i, given his side-information x_{K_i} and the broadcasted message. The side information of the receivers can be modeled by a directed graph $\mathcal{K}_n$, in which R_i observes the symbols $K_i := \{x_j : (i,j) \in E(\mathcal{K}_n)\}$. $\mathcal{K}_n$ is sometimes called the *knowledge graph*. A canonical example is where $\mathcal{K}_n$ is the complete graph (with no self-loops) on the vertex set $[n]$, i.e., each receiver observes all but his own symbol. In this simple case, broadcasting the sum $\sum_{i=1}^{n} x_i$ (in $\mathbb{F}$) allows each receiver to retrieve his own symbol, hence $m = 1$.

The problem was originally motivated by applications to distributed storage ([4], [5]), on-demand video streaming (ISCOD, [6]) and wireless networks (see, e.g., [7]), where a typical scenario is that clients miss information during transmissions of the network, and the network is interested in minimizing the retransmission length by exploiting the side-information clients already possess. More recently, Effros et al. [8] established a formal connection between index coding and the challenging problem of *network coding* ([9]), offering a partial explanation to the current barrier in understanding the capacity of general networks.

The minimum length of an index code for a given graph has well-known relations to other important graph parameters. For instance, it is bounded from below by the size of the maximum independent set, and it is bounded from above by the clique-cover number ($\chi(\bar{G})$) since for every clique in G, it suffices to broadcast a single symbol (recall the example above). The aforementioned connections also led to algorithmic connections (via convex relaxations) between the computational complexity of graph coloring and that of computing the minimum index code length of a graph ([10]). In theoretical computer science, index coding is related to some important communication models and problems in which players have overlapping information, such as the *one-way* communication complexity of the *index function* ([11]), and can also be viewed as an interesting special case of nondeterministic computation in the (notoriously difficult to understand) multiparty *Number-On-Forehead* communication model ([12]). We remark that index coding is also closely related to proving circuit lower bounds in complexity theory – Riis [13] observed that a certain index coding problem is equivalent to the so-called *shift conjecture* of Valiant [14] which, if true, would resolve a major open problem in complexity theory of proving superlinear lower bounds for logarithmic-depth circuits.

When the encoding function of the index code is *linear* in x (as in the example above), the corresponding scheme is called a *linear index code*. In their seminal paper, Bar-Yossef et al. [3] showed that the minimum length m of a *linear* index code is characterized precisely by a parameter of the knowledge graph $\mathcal{K}_n$, called the *minrank* ($\mathsf{minrk}_{\mathbb{F}}(\mathcal{K}_n)$), first introduced by Haemers [15] in the context of Shannon capacity of graphs.[1] Namely, $\mathsf{minrk}_{\mathbb{F}}(\mathcal{K}_n)$ is the minimum rank (over $\mathbb{F}$) of an $n \times n$ matrix M that "represents" $\mathcal{K}_n$. By "represents" we mean a matrix M that contains a zero in all entries corresponding to *non-edges*, and non-zero entries on the diagonal. Entries corresponding to edges are arbitrary. (Over $\mathbb{F}_2$ this is equivalent to being the adjacency matrix of a subgraph of $\mathcal{K}_n$, with diagonal entries set to one.) Note that without the "diagonal constraint", the above minimum would trivially be 0, and indeed this constraint is what makes the

A preliminary version of this paper ([1]) appeared in the proceedings of the 21st International Workshop on Randomization and Computation (RANDOM 2017). This research is supported by the Simons Collaboration on Algorithms and Geometry and by the National Science Foundation (NSF) under Grant No. CCF-1320188. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The work was done when Omri Weinstein was supported by a Simons Junior fellowship.

[1] To be precise, this holds only for graphs without self-loops. We will ignore this minor issue in this paper as it will not affect any of our results.

problem interesting and hard to analyze. While linear index codes are in fact optimal for a large class of knowledge graphs (including directed acyclic graphs, perfect graphs, odd "holes" and odd "anti-holes" [3]), there are examples where non-linear codes outperform their linear counterparts ([16]). In the same paper, Lubetzky and Stav [16] posed the following question about *typical* knowledge graphs, namely,

> *What is the minimum length of an index code for a random knowledge graph $\mathcal{K}_n = \mathcal{G}_{n,p}$?*

Here, $\mathcal{G}_{n,p}$ denotes a random Erdős-Rényi directed graph, i.e., a graph on n vertices in which each arc is taken independently with probability p. In this paper, we partially answer this open problem by determining the optimal length of *linear* index codes for such graphs. In other words, we prove a tight lower bound on the minrank of $\mathcal{G}_{n,p}$ for all values of $p \in [0, 1]$. In particular,

Theorem 1 (Main theorem, informal). *For any constant $0 < p < 1$ and any field $\mathbb{F}$ of cardinality $|\mathbb{F}| < n^{O(1)}$, it holds with high probability that*

$$\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) = \Theta\left(\frac{n}{\log n}\right) .$$

The formal quantitative statement of our result can be found in Corollary 2 below. We note that our general result (see Theorem 2) extends beyond the constant regime to *subconstant* values of p. Theorem 1 gives a near quadratic improvement over the previously best lower bound of $\Omega(\sqrt{n})$ ([16], [17]), and settles the linear index coding problem for random knowledge graphs, as an $O_p(n/\log n)$ linear index coding scheme is achievable via the clique-covering solution (see Section III-A).

A. Overview of the Proof of Theorem 1

In [16], Lubetzky and Stav showed that for any field $\mathbb{F}$ and a directed graph G,

$$\mathsf{minrk}_{\mathbb{F}}(G) \cdot \mathsf{minrk}_{\mathbb{F}}(\bar{G}) \geq n .$$

This inequality gives a lower bound of $\Omega(\sqrt{n})$ on the expected value of the minrank of $\mathcal{G}_{n,1/2}$. (Indeed, the random variables $\mathcal{G}_{n,1/2}$ and $\bar{\mathcal{G}}_{n,1/2}$ have identical distributions). Since $\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p})$ is monotonically non-increasing in p, the same bound holds for any $p \leq 1/2$. Haviv and Langberg [17] improved this result by proving a lower bound of $\Omega(\sqrt{n})$ for all constant p (and not just $p \leq 1/2$), and also by showing that the bound holds with high probability.

We now outline the main ideas of our proof. For simplicity we assume that $\mathbb{F} = \mathbb{F}_2$ and $p = 1/2$. To prove that $\mathsf{minrk}_2(\mathcal{G}_{n,p}) \geq k$, we need to show that with high probability, $\mathcal{G}_{n,p}$ has no representing matrix (in the sense of Definition 1 below) whose rank is less than k.

As a first attempt, we can show that any *fixed* matrix M with 1s on the diagonal of rank less than k has very low probability of representing a random graph in $\mathcal{G}_{n,p}$, and then apply a union bound over all such matrices M. Notice that this probability is simply 2^{-s+n}, where s is the sparsity of M (i.e., the number of non-zero entries) and the n is to account for the diagonal entries. Moreover, we observe that the sparsity s of any rank-k matrix with 1s on its main diagonal must be[2] at least $\approx n^2/k$. Finally, since the number of $n \times n$ matrices of rank k is $\approx 2^{2nk}$ (as a rank-k matrix can be written as a product of $n \times k$ by $k \times n$ matrices, which requires $2nk$ bits to specify), by a union bound, the probability that $\mathcal{G}_{n,p}$ contains a subgraph of rank $< k$ is bounded from above by (roughly) $2^{2nk} \cdot (1/2)^{n^2/k}$, which is $\ll 1$ for $k \leq \sqrt{n}/2$ (for large enough n). This recovers the previous $\Omega(\sqrt{n})$ lower bound of Haviv and Langberg [17] (for all constant p, albeit with a much weaker concentration bound).

To see why this argument is "stuck" at $\sqrt{n}$, we observe that we are not overcounting and indeed, there are $2^{n^{3/2}}$ matrices of rank $k \approx n^{1/2}$ and sparsity $s \approx n^{3/2}$. For instance, we can take the rank $n^{1/2}$ matrix that consists of $n^{1/2}$ diagonal $n^{1/2} \times n^{1/2}$ blocks of 1s (a disjoint union of $n^{1/2}$ equal-sized cliques), and replace the first $n^{1/2}$ columns with arbitrary values. Each such matrix has probability $2^{-n^{3/2}}$ of representing $\mathcal{G}_{n,p}$ (because of its sparsity) and there are $2^{n^{3/2}}$ of them, so the union bound breaks for $k = \Omega(\sqrt{n})$.

In order to go beyond $\sqrt{n}$, we need two main ideas. To illustrate the first idea, notice that in the above example, even though individually each matrix has probability $2^{-n^{3/2}}$ of representing $\mathcal{G}_{n,p}$, these "bad events" are highly correlated. In particular, each of these events implies that $\mathcal{G}_{n,p}$ must contain $n^{1/2} - 1$ disjoint cliques, an event that happens with roughly the same probability $2^{-n^{3/2}}$. Therefore, we see that the probability that the *union* of these bad events happens is only $2^{-n^{3/2}}$, greatly improving on the naive union bound argument. (We remark that this idea of "bunching together related events" is reminiscent of the chaining technique as used, e.g., in analyzing Gaussian processes.) More generally, the first idea (and also centerpiece) of our proof is Lemma 4, which shows that every matrix must contain a "nice" submatrix (in a sense to be defined below). The second and final idea, described in the next paragraph, will be to bound the number of "nice" submatrices, from which the proof would follow by a union bound over all such submatrices.

Before defining what we mean by "nice", we mention the following elementary yet crucial fact in our proof: Every rank k matrix is uniquely determined by specifying some k linearly independent rows, and some k linearly independent columns (i.e., a row space basis and a column space basis) including the indices of these rows and columns (see Lemma 2). This lemma implies that we can encode a matrix using only $\approx s_{basis} \cdot \log n$ bits, where s_{basis} is the minimal sparsity of a pair of row and column space bases that are guaranteed to exist. This in turn implies that there are only $\approx 2^{s_{basis} \log n}$ such matrices. Now, since the average number of 1s in a row or in a column of a matrix of sparsity s is s/n, one might hope that such a matrix contains a pair of row and column space bases of

[2] To see why, notice that any maximal linearly independent set of columns must "cover" all coordinates, i.e., there must not be any coordinate that is zero in all vectors, as otherwise we could take the column vector corresponding to that coordinate and it would be linearly independent of our set (due to the nonzero diagonal) in contradiction to maximality. Assuming all columns have roughly the same number of 1s, we obtain that each column has at least n/k 1s, leading to the claimed bound. See Lemma 3 for the full proof.

sparsity $k \cdot (s/n)$, and this is precisely our definition of a "nice" matrix. (Obviously, not all matrices are nice, and as the previous example shows, there are lots of "unbalanced" matrices where the nonzero entries are all concentrated on a small number of columns, hence they have no sparse column space basis even though the average sparsity of a column is very low; this is exactly why we need to go to submatrices.)

To complete this overview, notice that using the bound on the number of "nice" matrices, the union bound yields

$$2^{ks\log(n)/n} \cdot (1/2)^s ,$$

so one could set the rank parameter k to be as large as $\Theta(n/\log n)$ and the above expression would still be $\ll 1$. A similar bound holds for nice submatrices, completing the proof.

II. Preliminaries

For an integer n, we denote the set $\{1, \ldots, n\}$ by $[n]$. Throughout the paper by edges and arcs we mean undirected and directed edges, respectively. For an integer n and $0 \le p \le 1$, we denote by $\mathcal{G}_{n,p}$ the probability space over the directed graphs on n vertices where each arc is taken independently with probability p. By $\bar{G}$ we mean a directed graph on the same set of vertices as G that contains an arc if and only if G does not contain it.[3]

For a directed graph G, we denote by $\chi(G)$ the chromatic number of the undirected graph that has the same set of vertices as G, and an edge in place of every arc of G.

Let $\mathbb{F}$ be a finite field. For a vector $v \in \mathbb{F}^n$, we denote by v^j the jth entry of v, and by $v^{\le j} \in \mathbb{F}^j$ the vector v truncated to its first j coordinates. For a matrix $M \in \mathbb{F}^{n\times n}$ and indices $i, j \in [n]$, let $M_{i,j}$ be the entry in the ith row and jth column of M, $\mathrm{Col}_i(M)$ be the ith column of M, $\mathrm{Row}_i(M)$ be the ith row of M, and $\mathsf{rk}(M)$ be the rank of M over $\mathbb{F}$.

By a *principal submatrix* we mean a submatrix whose set of row indices is the same as the set of column indices. By the *leading principal submatrix* of size k we mean a principal submatrix that contains the first k columns and rows.

For a matrix $M \in \mathbb{F}^{n\times n}$, the sparsity $s(M)$ is the number of non-zero entries in M. We say that a matrix $M \in \mathbb{F}^{n\times n}$ of rank k *contains* an *s-sparse column (row) basis*, if M *contains* a column (row) basis (i.e., a set of k linearly independent columns (rows)) with a total of at most s non-zero entries. For a column (row) basis B of a matrix, its sparsity, denoted by $s(B)$, is the number of non-zero elements in B.

Definition 1 (Minrank [3], [16]). [4] *Let $G = (V, A)$ be a graph on $n = |V|$ vertices with the set of directed arcs A. A matrix $M \in \mathbb{F}^{n\times n}$* represents *$G$ if $M_{i,i} \neq 0$ for every $i \in [n]$, and $M_{i,j} = 0$ whenever $(i, j) \notin A$ and $i \neq j$. The minrank of G over $\mathbb{F}$ is*

$$\mathsf{minrk}_{\mathbb{F}}(G) = \min_{M \text{ represents } G} \mathsf{rk}(M) .$$

We say that two graphs *differ at only one vertex* if they differ only in arcs leaving one vertex. Following [18], [17], to amplify the probability in the main theorem, we shall use the following form of Azuma's inequality for the vertex exposure martingale.

Lemma 1 (Corollary 7.2.2 and Theorem 7.2.3 in [19]). *Let $f(\cdot)$ be a function that maps directed graphs to $\mathbb{R}$. If f satisfies the inequality $|f(H) - f(H')| \le 1$ whenever the graphs H and H' differ at only one vertex, then for any $\lambda > 0$,*

$$\Pr[|f(\mathcal{G}_{n,p}) - \mathbb{E}[f(\mathcal{G}_{n,p})]| > \lambda\sqrt{n-1}] < 2e^{-\lambda^2/2} .$$

III. The Minrank of a Random Graph

The following elementary linear-algebraic lemma shows that a matrix $M \in \mathbb{F}^{n\times n}$ of rank k is fully specified by k linearly independent rows, k linearly independent columns, and their $2k$ indices. In what follows, we denote by $\mathcal{M}_{n,k}$ the set of matrices from $\mathbb{F}^{n\times n}$ of rank k.

Lemma 2 (Row and column space bases encode the entire matrix). *The mapping $\phi\colon \mathcal{M}_{n,k} \to (\mathbb{F}^{1\times n})^k \times (\mathbb{F}^{n\times 1})^k \times [n]^{2k}$ defined as*

$$\phi(M) = (R, C, i_1, \ldots, i_k, j_1, \ldots, j_k) ,$$

is a one-to-one mapping, where $R = (\mathrm{Row}_{i_1}(M), \ldots, \mathrm{Row}_{i_k}(M))$ and $C = (\mathrm{Col}_{j_1}(M), \ldots, \mathrm{Col}_{j_k}(M))$ are, respectively, a row space basis and a column space basis of $M \in \mathcal{M}_{n,k}$ (taking, say, the lexicographically first if multiple bases exist).

Proof. We first claim that the intersection of R and C has full rank, i.e., that the submatrix $M' \in \mathbb{F}^{k\times k}$ obtained by taking rows $i_1, \ldots, i_k$ and columns $j_1, \ldots, j_k$ has rank k. This is a standard fact, see, e.g., [20, p20, Section 0.7.6]. We include a proof for completeness. Assume for convenience that $(i_1, \ldots, i_k) = (1, \ldots, k)$ and $(j_1, \ldots, j_k) = (1, \ldots, k)$. Next, assume towards contradiction that $\mathsf{rk}(M') = \mathsf{rk}(\{\mathrm{Col}_1(M'), \ldots, \mathrm{Col}_k(M')\}) = k' < k$. Since C is a column space basis of M, every column $\mathrm{Col}_i(M)$ is a linear combination of vectors from C, and in particular, every $\mathrm{Col}_i(M)^{\le k}$ is a linear combination of $\{\mathrm{Col}_1(M)^{\le k}, \ldots, \mathrm{Col}_k(M)^{\le k}\}$. Therefore, the $k \times n$ submatrix $M'' := (\mathrm{Col}_1^{\le k}(M), \ldots, \mathrm{Col}_n^{\le k}(M))$ has rank k'. On the other hand, the k rows of M'': $\mathrm{Row}_1(M), \ldots, \mathrm{Row}_k(M)$ were chosen to be linearly independent by construction. Thus, $\mathsf{rk}(M'') = k > k'$, which leads to a contradiction.

In order to show that ϕ is one-to-one, we show that R and C (together with their indices) uniquely determine the remaining entries of M. We again assume for convenience that $(i_1, \ldots, i_k) = (1, \ldots, k)$ and $(j_1, \ldots, j_k) = (1, \ldots, k)$. Consider any column vector $\mathrm{Col}_i(M)$, $i \in [n] \setminus [k]$. By definition, $\mathrm{Col}_i(M) = \sum_{t=1}^{k} \alpha_{i,t} \cdot \mathrm{Col}_t(M)$ for some coefficient vector $\alpha_i := (\alpha_{i,1}, \ldots, \alpha_{i,k}) \in \mathbb{F}^{k\times 1}$. Thus, in order to completely specify all the entries of $\mathrm{Col}_i(M)$, it suffices to determine the coefficient vector α_i. But M' has full rank, hence the equation

$$M' \alpha_i^T = \mathrm{Col}_i^{\le k}(M)$$

[3] Throughout the paper we assume that graphs under consideration do not contain self-loops. In particular, neither G nor $\bar{G}$ has self-loops.

[4] In this paper we consider the directed version of minrank. Since the minrank of a directed graph does not exceed the minrank of its undirected counterpart, a lower bound for a directed random graph implies the same lower bound for an undirected random graph. The bound is tight for both directed and undirected random graphs (see Theorem 3).

has a *unique* solution. Therefore, the coefficient vector α_i is fully determined by M' and $\mathrm{Col}_i^{\leq k}(M)$. Thus, the matrix M can be uniquely recovered from R, C and the indices $\{i_1, \ldots, i_k\}, \{j_1, \ldots, j_k\}$. □

The following corollary gives us an upper bound on the number of low-rank matrices that contain sparse column and row space bases. In what follows, we denote by $\mathcal{M}_{n,k,s}$ the set of matrices over $\mathbb{F}^{n\times n}$ of rank k that contain an s-sparse row space basis and an s-sparse column space basis.

Corollary 1 (Efficient encoding of sparse-base matrices)**.**

$$|\mathcal{M}_{n,k,s}| \leq (n \cdot |\mathbb{F}|)^{6s} \ .$$

Proof. Throughout the proof, we assume without loss of generality that $s \geq k$, as otherwise $|\mathcal{M}_{n,k,s}| = 0$ hence the inequality trivially holds. The function ϕ from Lemma 2 maps matrices from $\mathcal{M}_{n,k,s}$ to $(R, C, i_1, \ldots, i_k, j_1, \ldots, j_k)$, where R and C are s-sparse bases. Therefore, the total number of matrices in $\mathcal{M}_{n,k,s}$ is bounded from above by

$$\left(\binom{kn}{s} \cdot |\mathbb{F}|^s\right)^2 \cdot n^{2k} \leq \left((n^2)^s \cdot |\mathbb{F}|^s\right)^2 \cdot n^{2k} \leq (n \cdot |\mathbb{F}|)^{6s} \ ,$$

where the last inequality follows from $k \leq s$. □

Now we show that a matrix of low rank with nonzero entries on the main diagonal must contain many nonzero entries. To get some intuition on this, notice that a rank 1 matrix with nonzero entries on the diagonal must be nonzero everywhere. Also notice that the assumption on the diagonal is crucial – low rank matrices in general can be very sparse.

Lemma 3 (Sparsity vs. Rank for matrices with non-zero diagonal)**.** *For any matrix $M \in \mathbb{F}^{n\times n}$ with non-zero entries on the main diagonal (i.e., $M_{i,i} \neq 0$ for all $i \in [n]$), it holds that*

$$s(M) \geq \frac{n^2}{4\mathsf{rk}(M)} \ .$$

Proof. Let s denote $s(M)$. The average number of nonzero entries in a column of M is s/n. Therefore, Markov's inequality implies that there are at least $n/2$ columns in M *each of which* has sparsity at most $2s/n$. Assume without loss of generality that the first $n/2$ columns of M are such. Now pick a maximal set of linearly independent columns among these columns. We now finish the proof by showing that the cardinality of this set is at least $n^2/(4s)$. Indeed, in any set of less than $n^2/(4s)$ columns, the number of coordinates (i.e., row indices) that are nonzero in *at least one* of those columns is less than

$$\frac{n^2}{4s} \cdot \frac{2s}{n} = \frac{n}{2}$$

and therefore there exists a coordinate $i \in \{1, \ldots, n/2\}$ that is zero in all those columns. As a result, the ith column, which by assumption has a nonzero ith coordinate, must be linearly independent of all those columns, in contradiction to the maximality of the set. We therefore get that

$$\mathsf{rk}(M) \geq n^2/(4s) \ ,$$

as desired. □

The last lemma we need is also the least trivial. In order to use Corollary 1, we would like to show that any $n \times n$ matrix of rank k has sparse row and column space bases, where by sparse we mean that their sparsity is roughly k/n times that of the entire matrix. If the number of nonzero entries in each row and column was roughly the same, then this would be trivial, as we can take any maximal set of linearly independent columns or rows. However, in general, this might be impossible to achieve. E.g., consider the $n\times n$ matrix whose first k columns are chosen uniformly and the remaining $n-k$ columns are all zero. Then any column space basis would have to contain all first k columns (since they are linearly independent with high probability) and hence its sparsity is equal to that of the entire matrix. Instead, what the lemma shows is that one can always choose a *principal submatrix* with the desired property, i.e., that it contains sparse row and column space bases, while at the same time having relative rank that is at most that of the original matrix.

Lemma 4 (Every matrix contains a principal submatrix of low relative-rank and sparse bases)**.** *Let $M \in \mathcal{M}_{n,k}$ be a matrix. There exists a principal submatrix $M' \in \mathcal{M}_{n',k'}$ of M, such that $k'/n' \leq k/n$, and M' contains a column space basis and a row space basis of sparsity at most*

$$s(M') \cdot \frac{2k'}{n'} \ .$$

Note that if M contains a zero entry on the main diagonal, the lemma becomes trivial. Indeed, we can take M' to be a 1×1 principal submatrix formed by this zero entry. Thus, the lemma is only interesting for matrices M without zero elements on the main diagonal (i.e., when every principal submatrix has rank greater than 0).

Proof. We prove the statement of the lemma by induction on n. The base case $n = 1$ holds trivially.

Now let $n > 1$, and assume that the statement of the lemma is proven for every $m \times m$ matrix for $1 \leq m < n$. Let $s(i)$ be the number of nonzero entries in the ith column plus the number of non-zero entries in the ith row (note that a nonzero entry on the diagonal is counted twice). Let also $s_{\max} = \max_i s(i)$. By applying the same permutation to the columns and rows of M we can assume that $s(1) \leq s(2) \leq \cdots \leq s(n)$ holds.

If for some $1 \leq n' < n$, the leading principal submatrix M' of dimensions $n' \times n'$ has rank at most $k' \leq n'k/n$, then we use the induction hypothesis for M'. This gives us a principal submatrix M'' of dimensions $n'' \times n''$ and rank k'', such that M'' contains a column space basis and a row space basis of sparsity at most $s(M'') \cdot \frac{2k''}{n''}$. Also, by induction hypothesis $k''/n'' \leq k'/n' \leq k/n$, which proves the lemma statement in this case.

Now we assume that for all $n' < n$, the rank of the leading principal submatrix of dimension $n' \times n'$ is greater than $n'k/n$. We prove that the lemma statement holds for $M' = M$ for a column space basis, and an analogous proof gives the same result for a row space basis.

For every $0 \leq i \leq s_{\max}$, let $a_i = |\{j \ : \ s(j) = i\}|$. Note that

$$\sum_{i=0}^{s_{\max}} a_i = n \ . \qquad (1)$$

Let us select a column space basis of cardinality k by greedily adding linearly independent vectors to the basis in non-decreasing order of $s(i)$. Let k_i be the number of selected vectors j with $s(j) = i$. Then

$$\sum_{i=0}^{s_{\max}} k_i = k. \tag{2}$$

Next, for any $0 \le t < s_{\max}$, consider the leading principal submatrix given by indices i with $s(i) \le t$. The rank of this matrix is at most $k' = \sum_{i=0}^{t} k_i$, and its dimensions are $n' \times n'$, where $n' = \sum_{i=0}^{t} a_i < n$. Thus by our assumption $k'/n' \ge k/n$, or equivalently,

$$\sum_{i=0}^{t} k_i \ge \frac{k}{n} \cdot \sum_{i=0}^{t} a_i \ . \tag{3}$$

From (1) and (2),

$$\sum_{i=0}^{s_{\max}} k_i = \frac{k}{n} \cdot \sum_{i=0}^{s_{\max}} a_i \ . \tag{4}$$

Now, (3) and (4) imply that for all $0 \le t \le s_{\max}$:

$$\sum_{i=t}^{s_{\max}} k_i \le \frac{k}{n} \cdot \sum_{i=t}^{s_{\max}} a_i \ . \tag{5}$$

To finish the proof, notice that the sparsity of the constructed basis of M is at most

$$\begin{aligned}\sum_{i=1}^{s_{\max}} i \cdot k_i = \sum_{t=1}^{s_{\max}} \sum_{i=t}^{s_{\max}} k_i &\overset{(5)}{\le} \frac{k}{n} \cdot \sum_{t=1}^{s_{\max}} \sum_{i=t}^{s_{\max}} a_i \\ &= \frac{k}{n} \cdot \sum_{i=1}^{s_{\max}} i \cdot a_i = s(M) \cdot \frac{2k}{n} \ .\end{aligned}$$

□

Now we are ready to prove our main result – a lower bound on the minrank of a random graph.

Theorem 2.

$$\begin{aligned}\Pr\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) \ge \Omega\left(\frac{n\log(1/p)}{\log\left(n|\mathbb{F}|/p\right)}\right)\right] \\ \ge \ 1 - e^{-\Omega\left(\frac{n\log^2(1/p)}{\log^2(n|\mathbb{F}|/p)}\right)} \ .\end{aligned}$$

Proof. Let us bound from above probability that a random graph $\mathcal{G}_{n,p}$ has minrank at most

$$k := \frac{n\log(1/p)}{C\log\left(n|\mathbb{F}|/p\right)},$$

for some constant C to be chosen below.

Recall that by Lemma 4, every matrix of rank at most k contains a principal submatrix $M' \in \mathcal{M}_{n',k'}$ of sparsity $s' = s(M')$ with column and row space bases of sparsity at most

$$s' \cdot \frac{2k}{n},$$

where $k'/n' \le k/n$. By Corollary 1, there are at most $(n' \cdot |\mathbb{F}|)^{6(2s'k/n)}$ such matrices M', and (for any s') there are $\binom{n}{n'}$ ways to choose a principal submatrix of size n' in a matrix of size $n \times n$. Furthermore, recall that Lemma 3 asserts that for every n', k',

$$s' \ge \frac{n'^2}{4k'}. \tag{6}$$

Finally, since M' contains at least $s' - n'$ off-diagonal non-zero entries, M' represents $\mathcal{G}_{n',p}$ with probability at most $p^{s'-n'}$. We therefore have

$$\begin{aligned}&\Pr\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) \le k\right] \\ &\le \sum_{k',n',s'} \Pr\Big[\mathcal{G}_{n,p}\ M' \in \mathcal{M}_{n',k'} \text{ represents } \mathcal{G}_{n',p}, \\ &\qquad s(M') = s', s(\text{bases of } M') \le s' \cdot \frac{2k}{n}\Big] \\ &\le \sum_{k',n',s'} \binom{n}{n'} \cdot p^{s'-n'} \cdot (n' \cdot |\mathbb{F}|)^{12s'k/n} \\ &\le \sum_{k',n',s'} 2^{n'\log n - s'\log(1/p) + n'\log(1/p) + (12s'k/n)\log\left(n'|\mathbb{F}|\right)} \ ,\end{aligned} \tag{7}$$

where all the summations are taken over n', k', s.t. $k'/n' \le k/n$ and $s' \ge \frac{n'^2}{4k'}$, the first inequality is again by Lemma 4, and the second one is by Corollary 1. By $s(\text{bases of } M') \le s' \cdot \frac{2k}{n}$ we mean that M' contains row and column space bases of sparsity at most $s' \cdot \frac{2k}{n}$. We now argue that for sufficiently large constant C, all positive terms in the exponent of (7) are dominated by the magnitude of the negative term ($s'\log(1/p)$). Indeed,

$$\begin{aligned}&n'\log n + n'\log(1/p) + (12s'k/n)\log\left(n'|\mathbb{F}|\right) \\ &= n'\log\left(n/p\right) + (12s'k/n)\log\left(n'|\mathbb{F}|\right) \\ &\le (4s'k'/n')\log\left(n/p\right) + (12s'k/n)\log\left(n|\mathbb{F}|\right) \\ &\le (16s'k/n)\log\left(n|\mathbb{F}|/p\right) = (16s'/C)\log\left(1/p\right) \ ,\end{aligned}$$

where the first inequality follows from (6), and the second one follows from $k'/n' \le k/n$.

Since

$$\begin{aligned}s'\log(1/p) &\ge \frac{n'^2\log(1/p)}{4k'} \ge \frac{n\log(1/p)}{4k} \\ &= \frac{n\log(1/p)C\log\left(n|\mathbb{F}|/p\right)}{4n\log(1/p)} \ge \frac{C\log n}{4} \ ,\end{aligned}$$

we have that for every $C \ge 17$,

$$\begin{aligned}&\Pr\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) \le k = \frac{n\log(1/p)}{C\log\left(n|\mathbb{F}|/p\right)}\right] \\ &\le n^4 \cdot 2^{s'\log(1/p)\cdot(16/C-1)} \le n^4 \cdot 2^{-s'\log(1/p)/17} \\ &\le n^4 \cdot 2^{-C\log(n)/(17\cdot 4)} \le 0.5 \ .\end{aligned}$$

In particular, $\mathbb{E}\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p})\right] \ge \frac{n\log(1/p)}{2C\log\left(n|\mathbb{F}|/p\right)}$. Furthermore, note that changing a single row (or column) of a matrix can change its minrank by at most 1, hence the minrank of two graphs that differ in one vertex differs by at most 1. We may thus apply Lemma 1 with $\lambda = \Theta\left(\frac{\sqrt{n}\log(1/p)}{\log\left(n|\mathbb{F}|/p\right)}\right)$ to obtain

$$\begin{aligned}\Pr\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) \ge \Omega\left(\frac{n\log(1/p)}{\log\left(n|\mathbb{F}|/p\right)}\right)\right] \\ \ge \ 1 - e^{-\Omega\left(\frac{n\log^2(1/p)}{\log^2(n|\mathbb{F}|/p)}\right)} \ .\end{aligned}$$

as desired. □

Corollary 2. *For a constant $0 < p < 1$ and a field $\mathbb{F}$ of size $|\mathbb{F}| < n^{O(1)}$,*

$$\Pr\left[\, \mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) \geq \Omega(n/\log n) \,\right] \;\geq\; 1 - e^{-\Omega\left(n/\log^2 n\right)} \;.$$

A. Tightness of Theorem 2

In this section, we show that Theorem 2 provides a tight bound for all values of p bounded away from 1 (i.e., $p \leq 1 - \Omega(1)$). (See also the end of the section for the regime of p close to 1.)

Theorem 3. *For any p bounded away from 1,*

$$\Pr\left[\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p}) = O\left(\frac{n\log(1/p)}{\log n + \log(1/p)}\right)\right] \;\geq\; 1 - e^{-\Omega(n)} \;.$$

Proof. If $p \leq n^{-1/8}$, then $\frac{n\log(1/p)}{\log n + \log(1/p)} \geq \Omega(n)$, but $\mathsf{minrk}_{\mathbb{F}}(G)$ is always $\leq n$ which makes the statement trivial. Thus, in the following we assume that $p > n^{-1/8}$.

As we saw in the introduction, in the case of a clique (a graph with an arc between every pair of distinct vertices) it is enough to broadcast only one bit. This simple observation leads to the "clique-covering" upper bound: If a directed graph G can be covered by m cliques, then $\mathsf{minrk}_{\mathbb{F}}(G) \leq m$ ([21], [3], [17]). Note that the minimal number of cliques needed to cover G is exactly $\chi(\bar{G})$. Thus, we have the following upper bound: For any field $\mathbb{F}$ and any directed graph G,

$$\mathsf{minrk}_{\mathbb{F}}(G) \leq \chi(\bar{G}) \;. \tag{8}$$

Note that if we sample a graph from $\mathcal{G}_{n,p}$ and take its complement, the resulting graph is distributed according to $\mathcal{G}_{n,1-p}$. Now it follows from (8) that an upper bound on $\chi(\mathcal{G}_{n,1-p})$ with high probability, implies the same upper bound on $\mathsf{minrk}_{\mathbb{F}}(\mathcal{G}_{n,p})$.

Let $\mathcal{G}^{-}_{n,p}$ denote a random Erdős-Rényi *undirected* graph on n vertices, where each edge is drawn independently with probability p. For constant $0 < p < 1$, the classical result of Bollobás [22] asserts that the chromatic number of an undirected random graph satisfies

$$\Pr\left[\chi(\mathcal{G}^{-}_{n,1-p}) \leq \frac{n\log(1/p)}{2\log n}\left(1 + o(1)\right)\right] > 1 - e^{-\Omega(n)} \;. \tag{9}$$

In fact, Pudlák, Rödl, and Sgall [23] showed that (9) holds for any $p > n^{-1/4}$.

Since we define the chromatic number of a directed graph to be the chromatic number of its undirected counterpart, $\chi(\mathcal{G}_{n,1-p})$ and $\chi(\mathcal{G}^{-}_{n,1-p^2})$ have identical distributions. The bound (9) depends on p only logarithmically ($\log(1/p)$), thus, asymptotically the same bounds hold for the chromatic number of a random directed graph. □

The lower bound of Theorem 2 is also almost tight for the other extreme regime of $p = 1 - \varepsilon$, where $\varepsilon = o(1)$. Łuczak [24] proved that for $p = 1 - \Omega(1/n)$,

$$\Pr\left[\chi(\mathcal{G}^{-}_{n,1-p}) \leq \frac{n(1-p)}{2\log n(1-p)}\left(1 + o(1)\right)\right] \tag{10}$$
$$> 1 - \left(n(1-p)\right)^{-\Omega(1)} \;.$$

When $p = 1 - \varepsilon$, the upper bound (10) matches the lower bound of Theorem 2 for $\varepsilon \geq n^{-1+\Omega(1)}$. For $\varepsilon = O(n^{-1})$, (10) gives an asymptotically tight upper bound of $O(1)$. Thus, we only have a gap between the lower bound of Theorem 2 and known upper bounds when $p = 1 - \varepsilon$ and $\omega(1) \leq n\varepsilon \leq n^{o(1)}$.

Acknowledgements

We would like to thank Ishay Haviv for his valuable comments.

References

[1] A. Golovnev, O. Regev, and O. Weinstein, "The minrank of random graphs," in *RANDOM 2017*. Dagstuhl Publishing, 2017.

[2] Y. Birk and T. Kol, "Informed-source coding-on-demand (ISCOD) over broadcast channels," in *INFOCOM 1998*, 1998, pp. 1257–1264.

[3] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1479–1494, 2011.

[4] A. Mazumdar, "On a duality between recoverable distributed storage and index coding," in *ISIT 2014*. IEEE, 2014, pp. 1977–1981.

[5] F. Arbabjolfaei and Y. Kim, "Three stories on a two-sided coin: Index coding, locally recoverable distributed storage, and guessing games on graphs," in *Allerton Conf. Control, Communication and Computing 2015*. IEEE, 2015, pp. 843–850.

[6] Y. Birk and T. Kol, "Coding on demand by an informed source (ISCOD) for efficient broadcast of different supplemental data to caching clients," *IEEE Trans. Information Theory*, vol. 52, no. 6, pp. 2825–2830, 2006.

[7] R. W. Yeung and Z. Zhang, "Distributed source coding for satellite communications," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1111–1120, 1999.

[8] M. Effros, S. Y. E. Rouayheb, and M. Langberg, "An equivalence between network coding and index coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2478–2487, 2015.

[9] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[10] E. Chlamtac and I. Haviv, "Linear index coding via semidefinite programming," *Combinatorics, Probability & Computing*, vol. 23, no. 2, pp. 223–247, 2014.

[11] I. Kremer, N. Nisan, and D. Ron, "On randomized one-round communication complexity," *Computational Complexity*, vol. 8, no. 1, pp. 21–49, 1999.

[12] L. Babai, N. Nisan, and M. Szegedy, "Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs," *J. Comput. Syst. Sci.*, vol. 45, no. 2, pp. 204–232, 1992.

[13] S. Riis, "Information flows, graphs and their guessing numbers," *Electr. J. Comb.*, vol. 14, no. 1, 2007.

[14] L. G. Valiant, "Why is Boolean complexity theory difficult," *Boolean Function Complexity*, vol. 169, pp. 84–94, 1992.

[15] W. Haemers, "On some problems of Lovász concerning the Shannon capacity of a graph," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 231–232, 1979.

[16] E. Lubetzky and U. Stav, "Nonlinear index coding outperforming the linear optimum," *IEEE Trans. Inf. Theory*, vol. 8, no. 55, pp. 3544–3551, 2009.

[17] I. Haviv and M. Langberg, "On linear index coding for random graphs," in *ISIT 2012*. IEEE, 2012, pp. 2231–2235.

[18] H. T. Hall, L. Hogben, R. Martin, and B. Shader, "Expected values of parameters associated with the minimum rank of a graph," *Linear Algebra and its Applications*, vol. 433, no. 1, pp. 101–117, 2010.

[19] N. Alon and J. H. Spencer, *The Probabilistic Method*, ser. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2016.

[20] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed. Cambridge University Press, Cambridge, 2013.

[21] W. Haemers, "An upper bound for the Shannon capacity of a graph," in *Colloq. Math. Soc. János Bolyai*, vol. 25, 1978, pp. 267–272.

[22] B. Bollobás, "The chromatic number of random graphs," *Combinatorica*, vol. 8, no. 1, pp. 49–55, 1988.

[23] P. Pudlák, V. Rödl, and J. Sgall, "Boolean circuits, tensor ranks, and communication complexity," *SIAM J. Comput.*, vol. 26, no. 3, pp. 605–633, 1997.

[24] T. Łuczak, "The chromatic number of random graphs," *Combinatorica*, vol. 11, no. 1, pp. 45–54, 1991.