SN Operations Research Forum (2020) 1: 23
https://doi.org/10.1007/s43069-020-00020-5

ORIGINAL RESEARCH

Convex Optimization for the Densest Subgraph and Densest Submatrix Problems

Polina Bombina[1] · **Brendan Ames**[1]

Received: 12 March 2020 / Accepted: 23 July 2020 / Published online: 11 September 2020

Abstract

We consider the densest k-subgraph problem, which seeks to identify the k-node subgraph of a given input graph with maximum number of edges. This problem is well-known to be NP-hard, by reduction to the maximum clique problem. We propose a new convex relaxation for the densest k-subgraph problem, based on a nuclear norm relaxation of a low-rank plus sparse decomposition of the adjacency matrices of k-node subgraphs to partially address this intractability. We establish that the densest k-subgraph can be recovered with high probability from the optimal solution of this convex relaxation if the input graph is randomly sampled from a distribution of random graphs constructed to contain an especially dense k-node subgraph with high probability. Specifically, the relaxation is exact when the edges of the input graph are added independently at random, with edges within a particular k-node subgraph added with higher probability than other edges in the graph. We provide a sufficient condition on the size of this subgraph k and the expected density under which the optimal solution of the proposed relaxation recovers this k-node subgraph with high probability. Further, we propose a first-order method for solving this relaxation based on the alternating direction method of multipliers, and empirically confirm our predicted recovery thresholds using simulations involving randomly generated graphs, as well as graphs drawn from social and collaborative networks.

Keywords Densest subgraph · Submatrix localization · Nuclear norm relaxation · Maximum clique · Alternating direction method of multipliers

✉ Brendan Ames
bpames@ua.edu

Polina Bombina
pbombina@crimson.ua.edu

[1] Department of Mathematics, University of Alabama, Tuscaloosa, AL, USA

1 Introduction

We consider the *densest k-subgraph problem:* given graph $G = (V, E)$, identify the k-node subgraph of G of maximum density, i.e., maximum average degree. Equivalently, the problem reduces to finding the k-node subgraph of G with maximum number of edges. It is easy to see that the densest k-subgraph problem is NP-hard by reduction to the maximum clique problem, well-known to be NP-hard [1]. Indeed, if G contains a clique of size k, it would induce the densest k-subgraph of G; any polynomial-time algorithm for densest k-subgraph would immediately provide a polynomial-time algorithm for maximum clique. Moreover, it has been shown by [2–4] that the densest k-subgraph problem does not admit polynomial-time approximation schemes in general. Despite this intractability, the identification of dense subgraphs plays a significant role in many practical applications, especially in the analysis of web graphs, and social and biological networks [5–10].

We propose a new convex relaxation for the densest k-subgraph problem to address this intractability. Although we do not, and should not, expect our algorithm to provide a good approximation of the densest k-subgraph for all graphs, we will show that it is functionally equivalent to the densest k-subgraph problem for a large class of problem instances. In particular, suppose that the random input graph consists of a k-node subgraph H with edges added with significantly higher probability than those edges outside H. We will show that if k is sufficiently large then H is the densest k-subgraph of G, and it can be recovered from the optimal solution of our convex relaxation.

This result can be thought of as a specialization of recent developments regarding the recovery of clusters in graphs. In graph clustering, one seeks to partition the nodes of a given graph into dense subgraphs. Several recent results [11–28], among others, have established sufficient conditions on the generative model under which dense subgraphs can be recovered in a random graph, typically from the solution of some convex relaxation. These results assume that the random graph is generated using some generalization of the stochastic block model (see [29]), which assumes that the edges are added within blocks or clusters with higher frequency than between blocks, and provide sufficient conditions on the number and relative sizes of clusters, and the probabilities of adding edges that guarantee that the underlying block structure can be recovered in polynomial-time. The recent survey article [30] provides an overview of such recovery guarantees.

Relatively few analogous results exist for the densest k-subgraph problem. Ames and Vavasis [31, 32] consider convex relaxations for the maximum clique problem. Given an input graph, the *maximum clique problem* aims to identify the largest clique in the graph, that is, the vertex set of the largest complete subgraph (see [33, 34] for further discussion of the maximum clique problem). Ames and Vavasis [31, 32] establish that the maximum clique can be recovered from the optimal solution of particular convex relaxation if the input graph consists of a single large complete graph that is obscured by noise in the form of random edge additions and deletions. In particular, both results show that hidden cliques of size at least $\Omega(\sqrt{n})$ can be

identified with high probability for n-node random graphs constructed so that the probability of adding an edge between nodes in the hidden clique is significantly higher than adding other potential edges to the graph. The notation $f(x) = \Omega(g(x))$ indicates that there is some constant C such that $f(x) \geq Cg(x)$ for all sufficiently large x and we say that an event occurs *with high probability* if the probability of the event tends to 1 polynomially as the size of the graph N tends to ∞. The latter result recasts the hidden clique problem as that of finding the densest k-subgraph, where k is the size of the hidden clique. Similar theoretical recovery guarantees can be found in [35–40]. We delay the derivation of our convex relaxation and statement of the general recovery guarantee until Section 2.

We generalize these results for the densest subgraph problem to obtain an analogous recovery guarantee for the *densest submatrix problem*, which seeks to find the densest submatrix of given size. That is, we seek the submatrix of desired size with maximum number of nonzero entries. Similar results for the submatrix localization problem, where one seeks to find a block of entries with elevated mean in a random matrix, were presented in [41–43]. We will see that our convex relaxation correctly identifies the densest submatrix (of fixed size) in random matrices provided that entries within this submatrix are significantly more likely to be nonzero than an arbitrary entry of the matrix. We present our generalization of the densest subgraph problem to the densest submatrix problem and the statement of our theoretical recovery guarantees in Section 3. We provide proofs of our main results in Section 4 and conclude with discussion of a first-order method for solving our convex relaxations and empirical results illustrating efficacy of our approach in Section 5.

2 Relaxation of the Densest *k*-Subgraph Problem and Perfect Recovery of a Planted Clique

Our relaxation hinges on the observation, made in [32], that the adjacency matrix of *any* subgraph of G can be represented as the difference of a rank-one matrix and a binary correction matrix; this observation is closely related to the sparse plus low-rank decomposition of clustered graphs first considered in [25, 44], although with the restriction to submatrices of the adjacency matrix. Let $\hat{V} \subseteq V$ be a subset of nodes of $G = (V, E)$. We denote by $G(\hat{V})$ the subgraph induced by $\hat{V}$; that is, $G(\hat{V})$ is the graph with node set $\hat{V}$ and edge set given by the subset of E with both endpoints in $\hat{V}$. Let $\boldsymbol{v} \in \mathbf{R}^V$ be the characteristic vector of $\hat{V}$: $v_i = 1$ if $i \in \hat{V}$ and $v_i = 0$ otherwise for all $i \in V$. The matrix $\hat{\boldsymbol{X}} = \boldsymbol{v}\boldsymbol{v}^T$ is a rank-one binary matrix with nonzero entries indexed by $\hat{V} \times \hat{V}$. If $G(\hat{V})$ is a complete subgraph, i.e., $ij \in E$ for all $i, j \in \hat{V}$, then $\hat{\boldsymbol{X}}$ is equal to the sum of the adjacency matrix of $G(\hat{V})$ and the binary diagonal matrix $I_{\hat{V}}$ with nonzero entries indexed by $\hat{V}$; we call this sum the perturbed adjacency matrix of $G(\hat{V})$, and denote it by $\tilde{\boldsymbol{A}}_{\boldsymbol{G}(\hat{\boldsymbol{V}})}$.

If $G(\hat{V})$ is not a complete subgraph, then there is some $(i, j) \in V \times V, i \neq j$ such that $ij \notin E$. Let Ω denote the set of all such (i, j). For each $(i, j) \in \Omega$, we have

$\hat{X}_{ij} = 1$, while $[\tilde{A}_{G(\hat{V})}]_{ij} = 0$. It follows that $\tilde{A}_{G(\hat{V})} = \hat{X} - P_\Omega(\hat{X})$, where P_Ω is the projection onto the set of matrices having support contained in Ω, defined by

$$[P_\Omega(M)] = \begin{cases} M_{ij}, & \text{if}(i, j) \in \Omega \\ 0, & \text{otherwise.} \end{cases}$$

We call $(\hat{X}, \hat{Y}) := (\hat{X}, P_\Omega(\hat{X}))$ the *matrix representation* of the subgraph $G(\hat{V})$. The density of $G(\hat{V})$ is given by

$$d(G(\hat{V})) = \frac{1}{k}\left(\binom{k}{2} - \frac{1}{2}\|P_\Omega(\hat{X})\|_0\right),$$

where $\|M\|_0$ denotes the number of nonzero entries of M, because $\|P_\Omega(\hat{X})\|_0$ is equal to twice the number of nonadjacent nodes in $\hat{V}$. Moreover, the entries of the correction matrix $P_\Omega(\hat{X})$ are binary, which implies that $\|P_\Omega(\hat{X})\|_0 = \sum_{i,j\in V}[P_\Omega(\hat{X})]_{ij}$. Therefore, we may pose the densest k-subgraph problem as the rank-constrained binary program

$$\begin{aligned} &\min\ \mathrm{Tr}(Yee^T) \\ &\text{s.t.}\ \ \mathrm{Tr}(Xee^T) = k^2,\ P_\Omega(X - Y) = 0,\ \mathrm{rank}(X) = 1 \\ &\qquad X = X^T,\ Y = Y^T,\ X \in \{0,1\}^{V\times V},\ Y \in \{0,1\}^{V\times V}, \end{aligned} \tag{1}$$

where $\mathrm{Tr} : \mathbf{R}^{n\times n} \to \mathbf{R}^n$ denotes the matrix trace function, and e denotes the all-ones vector in $\mathbf{R}^V$. Unfortunately, combinatorial optimization problems involving rank and binary constraints are intractable in general. In particular, the densest k-subgraph problem is NP-hard and, hence, we cannot expect to be able to solve (1) efficiently. Relaxing the rank constraint with a nuclear norm penalty term given by $\|X\|_* = \sum_{i=1}^N \sigma_i(X)$ as in [45], the binary constraints with box constraints, and ignoring the symmetry constraints yields the convex program

$$\begin{aligned} &\min\ \|X\|_* + \gamma\mathrm{Tr}(Yee^T) \\ &\text{s.t.}\ \ \mathrm{Tr}(Xee^T) = k^2,\ P_\Omega(X - Y) = 0,\ 0 \le X \le ee^T,\ 0 \le Y, \end{aligned} \tag{2}$$

where $\gamma > 0$ is a regularization parameter chosen to control emphasis between the two objectives.

As mentioned earlier, we do not expect the solution of Eq. 2 to give a good approximation of the densest k-subgraph for an arbitrary graph. We instead restrict our focus to those graphs which we can expect to contain a single especially dense k-subgraph with high probability.

Definition 2.1 We construct the edge set of an N-node random graph $G = (V, E)$ as follows. Let $V^* \subseteq V$ be a k-subset of nodes; for each $(i, j) \in V^* \times V^*$, we add ij to E independently with probability q. For each $(i, j) \in (V \times V) - (V^* \times V^*)$, we add ij to E independently with probability $p < q$. We say such a graph G is sampled from the *planted dense k-subgraph model*.

This model, considered in [32], is a generalization of the planted clique model considered in [31], where q is chosen to be $q = 1$. On the other hand, the planted dense k-subgraph model is a special case of the generalized stochastic block model [44],

corresponding to a graph with exactly one cluster of size k and $N-k$ outlier nodes. Note that any graph G sampled from the planted dense k-subgraph contains a k-subgraph, $G(V^*)$, with higher density than the rest of the graph in expectation. Our goal is to derive conditions on the size k of this subgraph and the edge densities p and q that ensure recovery of the planted subgraph $G(V^*)$ from the optimal solution of Eq. 2. The following theorem provides such a sufficient condition.

Theorem 2.1 *Suppose that the N-node graph $G = (V, E)$ is sampled from the planted dense k-subgraph model with edge probabilities q and p respectively. Let $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ denote the matrix representation of the planted dense k-subgraph $G(V^*)$. Then, constants $c_1, c_2, c_3 > 0$ exist such that if*

$$q - p \geq c_1 \max\left\{\sqrt{\max\{\sigma_q^2, \sigma_p^2\}\frac{\log N}{k}}, \frac{\log N}{k}\sqrt{\sigma_p^2 N}, \frac{(\log N)^{3/2}}{k}\right\} \tag{3}$$

then $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ is the unique optimal solution of Eq. 2 *for penalty parameter*

$$\gamma \in \left(\frac{c_2}{(q-p)k}, \frac{c_3}{(q-p)k}\right), \tag{4}$$

and $G(V^*)$ is the unique densest k-subgraph of G with high probability; here σ_q^2 and σ_p^2 are equal to the edge creation variances $q(1-q)$ and $p(1-p)$ inside and outside of the planted dense k-subgraph, respectively.

Here, and in the rest of the paper, an event holding *with high probability (w.h.p.)* means that the event occurs with probability tending polynomially to one as $N \to \infty$; that is, there are scalars $\hat{c}_1, \hat{c}_2 > 0$ such that the event occurs with probability at least $1 - \hat{c}_1 N^{-\hat{c}_2}$. Note that Eq. 3 is only satisfiable when $k = \Omega((\log N)^{3/2})$. To illustrate the contribution of Theorem 2.1, we consider a few choices of p, q, and k.

First, suppose that p and q are fixed so that the edge densities in G are fixed as we vary N. In this case, Theorem 2.1 states that we may recover $G(V^*)$, with high probability, provided that $k = \Omega\left(\sqrt{N}\right)$. This bound is identical to that found many times in the planted clique literature [31, 35, 46–48], up to constants and the logarithmic term. It is widely believed that finding planted cliques of size $o(\sqrt{N})$ is intractable; indeed, several heuristic approaches have recently been proven to fail to recover planted cliques of size $o(\sqrt{N})$ in polynomial-time [2–4] and this intractability has been exploited in cryptographic applications [49].

On the other hand, our bound shows that planted cliques of size much smaller than $\sqrt{N}$ can be recovered in the presence of *sparse* noise. This should not be seen as a proof that we can recover planted cliques of size $o(\sqrt{N})$ in general, but rather evidence of the intimate relationship between the size of hidden cliques recoverable and the noise obscuring them. If very little noise in the form of diversionary edges is hiding the signal, here the planted clique, we should expect the signal to be significantly easier to recover. This is reflected in the fact that we can recover significantly smaller cliques than $o(\sqrt{N})$ in this setting. For example, let q be a fixed constant and let p vary with N such that $p \leq \log N/N$. The probability of adding an edge outside of $G(V^*)$ tends to zero as $k, N \to \infty$. Further, the left-hand side of Eq. 3

tends to q as $N \to \infty$, and the dominant term in the right-hand side is $(\log N)^{3/2}/k$. This implies that we can have exact recovery of the hidden clique w.h.p. provided $k = \Omega\left((\log N)^{3/2}\right)$. This lower bound on the size of recoverable k-subgraph matches that for identifying clusters in sparse graphs provided in the graph clustering literature, albeit for the case where the graph contains the single cluster indexed by V^* surrounded by many outlier nodes (see [30] and the references within). Moreover, this lower bound improves significantly upon that given by [32], where it is shown to that $k = \Omega(N^{1/3})$ is sufficient for exact recovery w.h.p. in the presence of sparse noise.

3 The Densest Submatrix Problem

The densest k-subgraph problem is a specialization of the far more general densest submatrix problem. Let $[M] = \{1, 2, \ldots, M\}$ for each positive integer M. Given a matrix $\boldsymbol{A} \in \mathbf{R}^{M\times N}$, the *densest $m\times n$-submatrix problem* seeks subsets $\bar{U} \subseteq [M]$ and $\bar{V} \subseteq [N]$ of cardinality $|\bar{U}| = m$ and $|\bar{V}| = n$, respectively, such that the submatrix $\boldsymbol{A}[\bar{U}, \bar{V}]$ with rows index by $\bar{U}$ and columns indexed by $\bar{V}$ contains the maximum number of nonzero entries. It should be clear that this specializes immediately to the densest k-subgraph problem when the input matrix is the perturbed adjacency matrix $\boldsymbol{A} = \boldsymbol{A}_G + \boldsymbol{I}$ of the input graph and $m = n = k$. However, the densest $m \times n$-submatrix problem allows far more flexible problem settings.

For example, the densest submatrix problem also specializes immediately to the maximum edge/density biclique problem. Let $G = (U, V, E)$ be a bipartite graph. Given integer m, n, the decision version of the maximum edge biclique problem determines if G contains an $m \times n$ biclique, i.e., whether there are vertex sets $\bar{U} \subseteq U$, $\bar{V} \subseteq V$ of cardinality $|\bar{U}| = m$ and $|\bar{V}| = n$ such that each vertex in $\bar{U}$ is adjacent to every vertex in $\bar{V}$. This problem immediately specializes to the densest $m \times n$-submatrix problem with $\boldsymbol{A}$ equal to the (U, V)-block of the adjacency matrix of G. Similar specializations exist for finding the densest subgraph in directed graphs, hypergraphs, and so on.

Let Ω denote the index set of zero entries of a given matrix $\boldsymbol{A} \in \mathbf{R}^{M\times N}$. Without loss of generality, we may assume that the entries of $\boldsymbol{A}$ are binary. If not, then we may replace $\boldsymbol{A}$ with the binary matrix having the same sparsity pattern without changing the index set of the densest $m \times n$-submatrix. We would like to obtain a rank-one matrix $\boldsymbol{X}$ with mn nonzero entries with minimum number of disagreements $\boldsymbol{A}$ on Ω:

$$\begin{array}{ll} \min\limits_{\boldsymbol{X},\boldsymbol{Y}\in\{0,1\}^{M\times N}} & \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{e}\boldsymbol{e}^T) \\ \text{s.t.} & \mathrm{Tr}(\boldsymbol{X}\boldsymbol{e}\boldsymbol{e}^T) = mn,\ P_\Omega(\boldsymbol{X} - \boldsymbol{Y}) = \boldsymbol{0},\ \mathrm{rank}\boldsymbol{X} = 1, \end{array} \tag{5}$$

where $\boldsymbol{Y}$ is used to count the number of disagreements between $\boldsymbol{A}$ and $\boldsymbol{X}$. Relaxing binary and rank constraints as before, we obtain the convex relaxation

$$\begin{array}{l} \min\ \|\boldsymbol{X}\|_* + \gamma\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{e}\boldsymbol{e}^T) \\ \text{s.t.}\ \ \mathrm{Tr}(\boldsymbol{X}\boldsymbol{e}\boldsymbol{e}^T) = mn,\ \ P_\Omega(\boldsymbol{X} - \boldsymbol{Y}) = \boldsymbol{0},\ \ \boldsymbol{0} \le \boldsymbol{X} \le \boldsymbol{e}\boldsymbol{e}^T,\ \ \boldsymbol{0} \le \boldsymbol{Y}, \end{array} \tag{6}$$

where $\gamma > 0$ is a regularization parameter chosen to tune between the two objectives. As before, we should expect to recover the solution of Eq. 5 from that of Eq. 6 when $\boldsymbol{A}$ contains a single large dense $m \times n$ block. The following definition proposes a class of random matrices with this property.

Definition 3.1 We construct an $M \times N$ random binary matrix $\boldsymbol{A}$ as follows. Let $U^* \subseteq [M]$ and $V^* \subseteq [N]$ be m and n-index sets. For each $i \in U^*$ and $j \in V^*$, we let $a_{ij} = 1$ with probability q and 0 otherwise. For each remaining (i, j), we set $a_{ij} = 1$ with probability $p < q$ and take $a_{ij} = 0$ otherwise. We say such a matrix $\boldsymbol{A}$ is sampled from the *planted dense $m \times n$-submatrix model.*

The following theorem provides a sufficient condition for exact recovery of a planted dense $m \times n$-submatrix generalizing the analogous result for recovery of a planted dense k-subgraph given by Theorem 2.1.

Theorem 3.1 *Suppose that the matrix $\boldsymbol{A} \in \mathbf{R}^{M \times N}$ is sampled from the planted dense $m \times n$-subgraph model with edge probabilities q and p, respectively, with rows and columns of the planted dense subgraph indexed by U^* and V^*, respectively. Let $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ denote the matrix representation of $\boldsymbol{A}(U^*, V^*)$. Let $N_{\max} := \max\{M, N\}$ and $n_{\min} := \min\{m, n\}$. Then, there are constants $c_1, c_2, c_3 > 0$ such that if*

$$q - p \geq c_1 \max\left\{\sqrt{\max\{\sigma_q^2, \sigma_p^2\}\frac{\log N_{\max}}{n_{\min}}}, \frac{\log N_{\max}}{n_{\min}}\sqrt{\sigma_p^2 N_{\max}}, \frac{(\log N_{\max})^{3/2}}{n_{\min}}\right\} \quad (7)$$

then $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ is the unique optimal solution of Eq. 6 *for penalty parameter $\gamma = t/((q - p)n_{\min})$ for all $c_2 \leq t \leq c_3$, and $\boldsymbol{A}(U^*, V^*)$ is the unique densest $m \times n$-submatrix of $\boldsymbol{A}$ with high probability.*

In the case when $M = N$ and $m = n$, the inequality Eq. 7 specializes to Eq. 3, although the constants c_1, c_2, c_3 should differ due to the lack of an assumption of symmetry of $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ in Theorem 3.1.

4 Derivation of the Recovery Guarantees

This section will consist of a proof of Theorem 3.1. The proof of Theorem 2.1 is identical except for minor modifications due to the symmetry of $\boldsymbol{A}$. We begin with the following theorem, which provides the required optimality conditions for Eq. 6.

Theorem 4.1 *Let $\bar{U} \subseteq \{1, \ldots, M\}$ be an m-subset of $[M]$ and let $\bar{V} \subseteq \{1, \ldots, N\}$ be an n-subset of $[N]$, and $\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}$ be their characteristic vectors. Then, the solutions*

$\bar{\boldsymbol{X}} = \bar{\boldsymbol{u}}\bar{\boldsymbol{v}}^T$ and $\bar{\boldsymbol{Y}} = P_\Omega(\bar{\boldsymbol{X}})$ are optimal for Eq. 6 *if and only if there are dual multipliers $\lambda \geq 0$, $\boldsymbol{\Lambda} \in \mathbf{R}_+^{M\times N}$, $\boldsymbol{\Xi} \in \mathbf{R}_+^{M\times N}$, and $\boldsymbol{W} \in \mathbf{R}^{M\times N}$ satisfying*

$$\frac{\bar{\boldsymbol{u}}\bar{\boldsymbol{v}}^T}{\sqrt{mn}} + \boldsymbol{W} - \lambda \boldsymbol{e}\boldsymbol{e}^T + \gamma \boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{\Xi} + \boldsymbol{\Lambda} = \boldsymbol{0} \tag{8a}$$

$$\mathrm{Tr}(\boldsymbol{\Lambda}^T(\bar{\boldsymbol{X}} - \boldsymbol{e}\boldsymbol{e}^T)) = 0 \tag{8b}$$

$$\mathrm{Tr}(\boldsymbol{\Xi}^T\bar{\boldsymbol{Y}}) = 0 \tag{8c}$$

$$\boldsymbol{W}^T\bar{\boldsymbol{u}} = \boldsymbol{0},\ \boldsymbol{W}\bar{\boldsymbol{v}} = \boldsymbol{0},\ \|\boldsymbol{W}\| \leq 1. \tag{8d}$$

The proof of Theorem 4.1 is nearly identical to that of [32, Theorem 4.1] and is omitted. Suppose that $\boldsymbol{A}$ is sampled from the planted dense (m, n)-subgraph model with edge probabilities $q > p$. Our goal is to establish the conditions on m, n, q, p given by Theorem 3.1 that guarantee exact recovery (w.h.p.) of the matrix representation $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$ of the planted submatrix with rows and columns given by $\bar{U}$ and $\bar{V}$ respectively. Our approach follows that of [32, Section 4]. We first explicitly construct dual multipliers $\boldsymbol{W}$ and $\boldsymbol{\Xi}$ using the duality feasibility condition given by Eq. 8a and the complementary slackness conditions given by Eq. 8b and Eq. 8c. We then use the characterization of the subdifferential of the nuclear norm given by Eq. 8d to construct the remaining dual variables $\lambda, \boldsymbol{\Lambda}$. We conclude the proof by using concentration inequalities to establish feasibility of the proposed dual variables under the hypothesis of Theorem 3.1.

We choose $\boldsymbol{W}$ and $\boldsymbol{\Xi}$ according to the dual feasibility condition given by Eq. 8a so that the orthogonality conditions $\boldsymbol{W}\bar{\boldsymbol{v}} = \boldsymbol{0}$ and $\boldsymbol{W}^T\bar{\boldsymbol{u}} = \boldsymbol{0}$ are satisfied. We consider the following cases.

Case 1. If $(i, j) \in \bar{U} \times \bar{V} - \Omega$, then Eq. 8a implies that

$$W_{ij} = \lambda - \frac{1}{\sqrt{mn}} - \Lambda_{ij} =: \tilde{\lambda} - \Lambda_{ij},$$

if we take $\Xi_{ij} = \gamma$ and define $\tilde{\lambda} := \lambda - 1/\sqrt{mn}$.
Case 2. If $i \in \bar{U}$, $j \in \bar{V}$, and $(i, j) \in \Omega$, then we have $\bar{X}_{ij} = \bar{Y}_{ij} = 1/\sqrt{mn}$, so $\Xi_{ij} = 0$ by Eq. 8c. It follows that $W_{ij} = \tilde{\lambda} - \gamma - \Lambda_{ij}$ in this case.
Case 3. If $(i, j) \notin \bar{U} \times \bar{V}$ such that $(i, j) \notin \Omega$ then we take $W_{ij} = \lambda$ and $\Xi_{ij} = \gamma$.
Case 4. If $i \notin \bar{U}$, $j \notin \bar{V}$ such that $(i, j) \in \Omega$, we take $W_{ij} = -\lambda p/(1 - p)$ and $\Xi_{ij} = \gamma - \lambda/(1 - p)$.
Case 5. If $i \in \bar{U}$ and $j \notin \bar{V}$ such that $(i, j) \in \Omega$, we take

$$W_{ij} = -\lambda\left(\frac{\nu_j}{m - \nu_j}\right),$$

where ν_j denotes the number of nonzero entries in the jth column of $\boldsymbol{A}$ indexed by rows in $\bar{U}$, so that $[\boldsymbol{W}^T\bar{\boldsymbol{u}}]_j = 0$. By our choice of W_{ij}, we have

$$\Xi_{ij} = \gamma - \frac{\lambda m}{m - \nu_j}.$$

Case 6. If $i \notin \bar{U}$, $j \in \bar{V}$, and $(i, j) \in \Omega$ then we take

$$W_{ij} = -\frac{\lambda \mu_i}{n - \mu_i} \Xi_{ij} = \gamma - \frac{\lambda n}{n - \mu_i},$$

where μ_i denotes the number nonzero entries in the ith row of $\boldsymbol{A}$ indexed by columns in $\bar{V}$.

By our choice of $\boldsymbol{W}$ and $\boldsymbol{\Xi}$, we have $[\boldsymbol{W}\bar{\boldsymbol{v}}]_i = 0$ for all $i \notin \bar{U}$ and $[\boldsymbol{W}^T\bar{\boldsymbol{u}}]_i = 0$ for all $i \notin \bar{V}$. We choose the remaining dual variables λ and $\boldsymbol{\Lambda}$ so that $[\boldsymbol{W}\bar{\boldsymbol{v}}]_i = 0$ for all $i \in \bar{U}$ and $[\boldsymbol{W}^T\bar{\boldsymbol{u}}]_i = 0$ for all $i \in \bar{V}$.

The orthogonality conditions $\boldsymbol{W}^T\bar{\boldsymbol{u}} = \boldsymbol{0}$ and $\boldsymbol{W}\bar{\boldsymbol{v}} = \boldsymbol{0}$ define a linear system with $m + n$ equations for the mn unknown entries of $\boldsymbol{\Lambda}$ when all other dual variables are fixed. To obtain a particular solution of this underdetermined linear system, we make the additional assumption that $\boldsymbol{\Lambda}(\bar{U}, \bar{V})$ has rank at most 2, taking the form $\boldsymbol{\Lambda}(\bar{U}, \bar{V}) = \boldsymbol{y}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{z}^T$ for some $\boldsymbol{y} \in \mathbf{R}^m$ and $\boldsymbol{z} \in \mathbf{R}^n$. Under this assumption, the conditions $[\boldsymbol{W}\boldsymbol{v}]_i = 0$, $i \in \bar{U}$ and $[\boldsymbol{W}^T\boldsymbol{u}]_j = 0$, $j \in \bar{V}$ yield the linear system

$$\begin{pmatrix} n\boldsymbol{I} & \boldsymbol{e}\boldsymbol{e}^T \\ \boldsymbol{e}\boldsymbol{e}^T & m\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} -\gamma\bar{\boldsymbol{\mu}} + n\tilde{\lambda}\boldsymbol{e} \\ -\gamma\bar{\boldsymbol{v}} + m\tilde{\lambda}\boldsymbol{e} \end{pmatrix}, \tag{9}$$

where the vectors $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{v}}$ are defined by $\bar{\mu}_i = n - \mu_i$ for all $i \in \bar{U}$ and $\bar{v}_j = m - v_j$ for all $j \in \bar{V}$. It is easy to see that this system is singular with null space spanned by $(\boldsymbol{e}; -\boldsymbol{e})$. However, it is also easy to see that the unique solution of

$$\begin{pmatrix} n\boldsymbol{I} + \boldsymbol{e}\boldsymbol{e}^T & \boldsymbol{0} \\ \boldsymbol{0} & m\boldsymbol{I} + \boldsymbol{e}\boldsymbol{e}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} = \begin{pmatrix} -\gamma\bar{\boldsymbol{\mu}} + n\tilde{\lambda}\boldsymbol{e} \\ -\gamma\bar{\boldsymbol{v}} + m\tilde{\lambda}\boldsymbol{e} \end{pmatrix} \tag{10}$$

is a solution of Eq. 9; see [32, Section 4.2] for further details. Applying the Sherman-Morrison-Woodbury formula (see [50, Equation (2.1.4)]), we have

$$\boldsymbol{y} = \frac{1}{n}\left(\tilde{\lambda}\frac{n^2}{m+n} - \gamma\bar{\boldsymbol{\mu}} + \gamma\frac{\bar{\boldsymbol{\mu}}^T\boldsymbol{e}}{m+n}\boldsymbol{e}\right), \quad \boldsymbol{z} = \frac{1}{m}\left(\tilde{\lambda}\frac{m^2}{m+n} - \gamma\bar{\boldsymbol{v}} + \gamma\frac{\bar{\boldsymbol{v}}^T\boldsymbol{e}}{m+n}\boldsymbol{e}\right). \tag{11}$$

The entries of $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{v}}$ are binomial random variables corresponding to n and m independent Bernoulli trials ith probability of success $1 - q$, respectively. Therefore, we have

$$\mathbb{E}[\boldsymbol{y}] = \frac{n}{m+n}\left(\tilde{\lambda} - \gamma(1-q)\right)\boldsymbol{e}, \quad \mathbb{E}[\boldsymbol{z}] = \frac{m}{m+n}\left(\tilde{\lambda} - \gamma(1-q)\right)\boldsymbol{e}. \tag{12}$$

Choosing $\lambda = \frac{1}{\sqrt{mn}} + \gamma(1-q) + \gamma\tau$ for some $\tau > 0$ to be chosen later ensures that the entries of $\boldsymbol{\Lambda}$ are strictly positive in expectation.

We next describe how to choose τ so that the entries of $\boldsymbol{y}$ and $\boldsymbol{z}$ are positive with high probability. To do so, we will make repeated use of the following specialization of the classical Bernstein inequality to bound the sum of independent Bernoulli random variables (see, for example, [51, Section 2.8]).

Lemma 4.1 *Let $x_1, \ldots, x_k$ be a sequence of k independent $\{0, 1\}$ Bernoulli random variables, each with probability of success ρ. Let $s = \sum_{i=1}^{k} x_i$ be the binomially distributed random variable denoting the number of successes. Then,*

$$\Pr\left(|s - \rho k| > 6\max\left\{\sqrt{\rho(1-\rho)k\log t}, \log t\right\}\right) \leq 2t^{-6}. \tag{13}$$

Applying Lemma 4.1 with $t = N$ to each component of $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\nu}}$ and the union bound shows that

$$|\bar{\mu}_i - (1-q)n| \leq 6\max\{\sqrt{\sigma_q^2 n \log N}, \log N\} \tag{14}$$

$$|\bar{\nu}_j - (1-q)m| \leq 6\max\{\sqrt{\sigma_q^2 m \log N}, \log N\} \tag{15}$$

for all $i \in \bar{U}$ and $j \in \bar{V}$ w.h.p., where $\sigma_q^2 = q(1-q)$. On the other hand, $\bar{\boldsymbol{\nu}}^T \boldsymbol{e} = \bar{\boldsymbol{\mu}}^T \boldsymbol{e}$ is equal to the number of nonzero entries in the $\bar{U} \times \bar{V}$ block of $\boldsymbol{A}$. Therefore, $\bar{\boldsymbol{\nu}}^T \boldsymbol{e} = \bar{\boldsymbol{\mu}}^T \boldsymbol{e}$ is a binomially distributed random variable, with $\mathbb{E}[\bar{\boldsymbol{\nu}}^T \boldsymbol{e}] = \mathbb{E}[\bar{\boldsymbol{\mu}}^T \boldsymbol{e}] = mn(1-q)$. Applying Lemma 4.1 with $t = N$ again establishes that

$$|\bar{\boldsymbol{\nu}}^T \boldsymbol{e} - (1-q)mn| = |\bar{\boldsymbol{\mu}}^T \boldsymbol{e} - (1-q)mn| \leq 6\max\{\sqrt{\sigma_q^2 mn \log N}, \log N\} \tag{16}$$

w.h.p. It follows immediately that

$$\begin{aligned} |y_i - \mathbb{E}[y_i]| &\leq \frac{\gamma}{n}\left(|\bar{\mu}_i - \mathbb{E}[\bar{\mu}_i]| + \frac{1}{m+n}|\bar{\boldsymbol{\mu}}^T \boldsymbol{e} - \mathbb{E}[\bar{\boldsymbol{\mu}}^T \boldsymbol{e}]|\right) \\ &\leq 6\gamma\left(1 + \frac{1}{\sqrt{m}}\right)\max\left\{\sqrt{\sigma_q^2 \frac{\log N}{n}}, \frac{\log N}{n}\right\} \end{aligned} \tag{17}$$

for each $i \in \bar{U}$ if Eqs. 14 and 16 are satisfied. Following an identical argument, we see that

$$|z_i - \mathbb{E}[z_i]| \leq 6\gamma\left(1 + \frac{1}{\sqrt{n}}\right)\max\left\{\sqrt{\sigma_q^2 \frac{\log N}{m}}, \frac{\log N}{m}\right\} \tag{18}$$

if Eqs. 15 and 16 hold. Substituting Eqs. 17 and 18 into the formula for Λ_{ij} shows that

$$\begin{aligned} \Lambda_{ij} &= y_i + z_j \geq \mathbb{E}[y_i] - |y_i - \mathbb{E}[y_i]| + \mathbb{E}[z_j] - |z_j - \mathbb{E}[z_j]| \\ &\geq \gamma\tau - 12\gamma\left(1 + \frac{1}{\sqrt{m}}\right)\max\left\{\sqrt{\sigma_q^2 \frac{\log N}{m}}, \frac{\log N}{m}\right\} \end{aligned}$$

for all $i \in \bar{U}$, $j \in \bar{V}$ w.h.p.; here, we use the assumption that $m \leq n$. Choosing

$$\tau = 12\left(1 + \frac{1}{\sqrt{m}}\right)\max\left\{\sqrt{\sigma_q^2 \frac{\log N}{m}}, \frac{\log N}{m}\right\} \tag{19}$$

ensures that the entries of $\boldsymbol{\Lambda}$ are nonnegative w.h.p.

4.1 Nonnegativity of $\boldsymbol{\Xi}$

We next establish conditions on the regularization parameter γ ensuring that the entries of the dual variable $\boldsymbol{\Xi}$ are nonnegative. Recall that Ξ_{ij} takes value 0 or γ for all (i, j) except those corresponding to Cases 4 through 6 in the choice of $\boldsymbol{W}$ and $\boldsymbol{\Xi}$.

We begin with Case 5 in the construction of $\boldsymbol{W}$ and $\boldsymbol{\Xi}$. Recall that

$$\Xi_{ij} = \gamma - \frac{\lambda m}{m - \nu_j} = \frac{1}{m - \nu_j}\left(\gamma(mq - \nu_j) - \gamma m\tau - \frac{m}{\sqrt{mn}}\right) \tag{20}$$

if $i \in \bar{U}$ and $j \notin \bar{V}$ such that $ij \in \Omega$. Since ν_j is a binomial random variable corresponding to m independent Bernoulli trials with probability of success p, applying Bernsteins inequality, given by Eq. 13 shows that $\nu_j \geq pm + 6\max\{\sqrt{\sigma_p^2 m \log N}, \log N\}$, and, hence,

$$\Xi_{ij} \geq \frac{m}{m - \nu_j}\left(\gamma\left(q - p - 6\max\left\{\sqrt{\sigma_p^2\frac{\log N}{m}}, \frac{\log N}{m}\right\} - \tau\right) - \frac{1}{\sqrt{mn}}\right)$$

w.h.p., where $\sigma_p^2 := p(1 - p)$. Under the gap assumption

$$q - p \geq 18\left(1 + \frac{1}{\sqrt{m}}\right)\max\left\{\max\{\sigma_q, \sigma_p\}\sqrt{\frac{\log N}{m}}, \frac{\log N}{m}\right\} \tag{21}$$

and the choice of τ given by Eq. 19, we see that

$$\Xi_{ij} \geq \frac{m}{m - \nu_j}\left(\frac{\gamma}{\tilde{c}}(q - p) - \frac{1}{\sqrt{mn}}\right)$$

w.h.p. for some constant $\tilde{c} \geq 3$. An identical bound holds for entries of $\boldsymbol{\Xi}$ corresponding to Case 6 by symmetry. Finally, the bound for Case 4 follows by substituting $\nu_i = pm$ in Eq. 20 which establishes that $\Xi_{ij} \geq 0$ if $\gamma(q - p) \geq 3/\sqrt{mn}$ in this case. Applying the union bound over all entries in $\boldsymbol{\Xi}$ establishes that $\boldsymbol{\Xi}$ is nonnegative w.h.p. if q and p satisfy the gap assumption given by Eq. 21 and

$$\gamma \geq \frac{\tilde{c}}{(q - p)\sqrt{mn}}. \tag{22}$$

4.2 A Bound on the Matrix W

To complete the proof, we derive a sufficient condition involving m, n, M, N, p, and q that ensures that $\boldsymbol{W}$, as constructed above, satisfies $\|\boldsymbol{W}\| < 1$ with high probability. To simplify our notation, we again make the assumption that $m \leq n$ and $M \leq N$. Our analysis will translate superficially to the cases when $m \leq n$ and $M \geq N$, $m \geq n$ and $M \leq N$, and $m \geq n$ and $M \geq N$. We bound $\|\boldsymbol{W}\|$ using the triangle inequality and the decomposition $\boldsymbol{W} = \gamma\boldsymbol{Q} + \lambda\boldsymbol{S}$, where $\gamma Q_{ij} = W_{ij}$ if $i \in \bar{U}$, $j \in \bar{V}$ and $\gamma Q_{ij} = 0$ otherwise. To bound the norms of $\boldsymbol{Q}$ and $\boldsymbol{S}$, we will make repeated use of the following bound on the norm of a random matrix. Specifically, Lemma 4.2 is a special case of the matrix concentration inequality given by [52, Corollary 3.11] on the spectral norm of matrices with i.i.d. mean zero bounded entries.

Lemma 4.2 *Let $\boldsymbol{A} = [a_{ij}] \in \mathbf{R}^{m\times n}$ be a random matrix with i.i.d. mean zero entries a_{ij} having variance σ^2 and satisfying $|a_{ij}| \le B$. Let $n_{\max} = \max\{m, n\}$. Then, there is a constant $c > 0$ such that*

$$\Pr\left(\|\boldsymbol{A}\| > c\max\left\{\sqrt{\sigma^2 n_{\max}}, \sqrt{B\log t}\right\}\right) \le n_{\max} t^{-7} \tag{23}$$

for all $t > 0$.

The following lemma provides the desired bound on $\|\boldsymbol{Q}\|$.

Lemma 4.3 *Suppose that the matrix $\boldsymbol{W}$ is constructed according to Cases 1 through 6 for a matrix $\boldsymbol{A}$ sampled from the planted dense submatrix model with $m \le n$ and $M \le N$. Then, there is a constant $C_Q > 0$ such that*

$$\|\boldsymbol{Q}\| \le C_Q \max\left\{\sqrt{\sigma_q^2 n \log N}, \sqrt{\frac{n}{m}}\log N\right\}$$

with high probability.

We delay the proof of Lemma 4.3 until Appendix A. The following lemma provides the required bound on $\|\boldsymbol{S}\|$. Our analysis follows a similar argument to that of [31, Section 4.2]; we include it here for completeness.

Lemma 4.4 *Suppose that the matrix $\boldsymbol{W}$ is constructed according to Cases 1 through 6 for a matrix $\boldsymbol{A}$ sampled from the planted dense submatrix model with $m \le n$ and $M \le N$. Let $\tilde{\sigma}_p^2 := p/(1-p)$ and $B := \max\{1, \tilde{\sigma}_p^2\}$. Assume that p is bounded away from 1 so that $B = O(1)$. Then, exists constant $C_S > 0$ such that*

$$\|\boldsymbol{S}\| \le C_S \max\left\{\sqrt{\tilde{\sigma}_p^2 N \log N}, \log N\right\}\sqrt{\log N}$$

with high probability.

It follows immediately from Lemmas 4.3 and 4.4 that

$$\|\boldsymbol{W}\| \le C_Q\gamma\max\left\{\sqrt{\sigma_q^2 n\log N}, \sqrt{\frac{m}{n}}\log N\right\} + C_S\lambda\max\left\{\sqrt{\tilde{\sigma}_p^2 N}\log N, (\log N)^{3/2}\right\} \tag{24}$$

w.h.p. On the other hand,

$$\begin{aligned}\lambda &= \frac{1}{\sqrt{mn}} + \gamma(1-q) + \gamma\tau \\ &\le \frac{1}{\sqrt{mn}}\left(1 + \frac{\tilde{c}}{(q-p)\sqrt{mn}}1 - q + \frac{1}{2}(q-p)\right) \le \frac{\tilde{c}}{(q-p)\sqrt{mn}}(1-p),\end{aligned}$$

where we obtain the first inequality by substituting the choice of γ given by Eq. 22 and the upper bound $\tau \le 3(q-p)/2 \le \tilde{c}(q-p)/2$. We obtain the last inequality using the fact that $(1/\tilde{c} + 1/2)(q-p) \le q - p$. Further, we have

$$\lambda\tilde{\sigma}_p \le \frac{\tilde{c}(1-p)}{(q-p)\sqrt{mn}}\sqrt{\frac{p}{1-p}} = \frac{\tilde{c}\sqrt{p(1-p)}}{(q-p)\sqrt{mn}} = \frac{\tilde{c}\sigma_p}{(q-p)\sqrt{mn}}.$$

Substituting back into Eq. 24, we see that

$$\begin{aligned}\|\boldsymbol{W}\| &= O\left(\frac{1}{(q-p)\sqrt{mn}}\left(\max\left\{\sqrt{\sigma_q^2 n\log N}, \sqrt{\frac{n}{m}}\log N\right\} + \max\left\{\sqrt{\sigma_p^2 N}\log N, (\log N)^{3/2}\right\}\right)\right)\\ &= O\left(\frac{1}{q-p}\max\left\{\sqrt{\frac{\sigma_q^2\log N}{m}}, \sqrt{\frac{\sigma_p^2 N}{mn}}\log N, \frac{(\log N)^{3/2}}{m}\right\}\right) \qquad (25)\end{aligned}$$

w.h.p. Enforcing $q - p$ so that Eq. 21 holds and the right-hand side of Eq. 25 is bounded above by 1 establishes Theorem 3.1. This completes the proof.

5 A First-order Method Based on the Alternating Direction Method of Multipliers

We conclude with discussion of an optimization algorithm for solution of Eq. 6 based on the alternating direction method of multipliers (ADMM); see [53] for details regarding the ADMM. We first present a derivation of the method and then empirically validate its performance using randomly generated matrices and real-world collaboration and communication networks.

5.1 The Optimization Algorithm

To apply the ADMM to Eq. 6, we first introduce artificial variables $\boldsymbol{Q}$, $\boldsymbol{W}$, $\boldsymbol{Z}$ to obtain the equivalent convex optimization problem

$$\begin{aligned}\min\ &\|\boldsymbol{X}\|_* + \boldsymbol{\gamma}\|\boldsymbol{Y}\|_1 + \mathbb{1}_{\Omega_Q}(\boldsymbol{Q}) + \mathbb{1}_{\Omega_W}(\boldsymbol{W}) + \mathbb{1}_{\Omega_Z}(\boldsymbol{Z})\\ &\boldsymbol{X} = \boldsymbol{Y} = \boldsymbol{Q},\ \boldsymbol{X} - \boldsymbol{W} = \boldsymbol{0},\ \boldsymbol{X} - \boldsymbol{Z} = \boldsymbol{0},\ \boldsymbol{Y} \geq \boldsymbol{0},\end{aligned} \qquad (26)$$

where Ω_Q, Ω_W, Ω_Z denote the constraint sets

$$\Omega_Q := \{\boldsymbol{Q} : P_\Omega(\boldsymbol{Q}) = \boldsymbol{0}\},\ \Omega_W := \{\boldsymbol{W} : \boldsymbol{e}^T\boldsymbol{W}\boldsymbol{e} = mn\},\ \Omega_{Z:=} = \{\boldsymbol{Z} : Z_{ij} \leq 1\ \forall (i,j) \in M \times N\},$$

where Ω indicates the set of $(i, j) \in V \times V$ such that $a_{ij} = 0$, and P_Ω denotes the projection onto the set of all matrices having support contained in Ω. Here, $\mathbb{1}_S : \mathbf{R}^{M\times M} \to \{0, +\infty\}$ is the indicator function of the set $S \subseteq \mathbf{R}^{M\times N}$, such that $\mathbb{1}_S(\boldsymbol{X}) = 0$ if $\boldsymbol{X} \in S$, and $+\infty$ otherwise. We solve Eq. 26 iteratively using the ADMM. Specifically, we update each primal variable by minimizing the augmented Lagrangian in Gauss-Seidel fashion with respect to each primal variable. Then, the dual variables are updated using the updated primal variables. The augmented Lagrangian of Eq. 26 is given by

$$\begin{aligned}L_\tau = \|\boldsymbol{X}\|_* &+ \boldsymbol{\gamma}\|\boldsymbol{Y}\|_1 + \mathbb{1}_{\Omega_Q}(\boldsymbol{Q}) + \mathbb{1}_{\Omega_W}(\boldsymbol{W}) + \mathbb{1}_{\Omega_Z}(\boldsymbol{Z}) + \mathrm{Tr}(\boldsymbol{\Lambda}_Q(\boldsymbol{X} - \boldsymbol{Y} - \boldsymbol{Q})) + \mathrm{Tr}(\boldsymbol{\Lambda}_W(\boldsymbol{X} - \boldsymbol{W}))\\ &+ \mathrm{Tr}(\boldsymbol{\Lambda}_Z(\boldsymbol{X} - \boldsymbol{Z})) + \frac{\tau}{2}\left(\|\boldsymbol{X} - \boldsymbol{Y} - \boldsymbol{Q}\|_F^2 + \|\boldsymbol{X} - \boldsymbol{W}\|_F^2 + \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2\right),\end{aligned}$$

where τ is a regularization parameter chosen so that L_τ is strongly convex in each primal variable. Minimization of the augmented Lagrangian with respect to each of the artificial primal variables $\boldsymbol{Q}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ is equivalent to projection onto each of the sets Ω_Q, Ω_W, and Ω_Z; each of these projections has an analytic expression.

We update $\boldsymbol{Y}$ using projection onto the nonnegative cone: $P_{\mathbf{R}^{M\times N}_{+}}(\boldsymbol{M})$ is the matrix with ijth entry m_{ij} if $m_{ij} \geq 0$ and 0 otherwise. On the other hand, we update $\boldsymbol{X}$ using the proximal function for the nuclear norm $\|\cdot\|_*$, which can be computed by applying the soft thresholding operator defined by $S_\phi(\boldsymbol{x}) = \text{sign}(\boldsymbol{x})\max\{|\boldsymbol{x}| - \phi\boldsymbol{e}, \boldsymbol{0}\}$ to the vector of singular values. Here, $\text{sign}(\boldsymbol{x})$ is the vector whose entries are the signs of the corresponding entries of $\boldsymbol{x}$, $|\boldsymbol{x}|$ denotes the vector whose entries are the magnitudes of the corresponding entries of $\boldsymbol{x}$, and the maximum denotes the vector of pairwise maximums. We declare the algorithm to have converged when the primal and dual residuals $\|\boldsymbol{X}^{l+1} - \boldsymbol{W}^{l+1}\|_F$, $\|\boldsymbol{X}^{l+1} - \boldsymbol{Z}^{l+1}\|_F$, $\|\boldsymbol{W}^{l+1} - \boldsymbol{W}^{l}\|_F$, $\|\boldsymbol{Z}^{l+1} - \boldsymbol{Z}^{l}\|_F$, and $\|\boldsymbol{Q}^{l+1} - \boldsymbol{Q}^{l}\|_F$ are smaller than a desired error tolerance. The steps of the algorithm are summarized in Algorithm 1.[1]

Algorithm 1 ADMM for solving relaxation of densest (m, n)-submatrix problem (6).

Inputs Binary matrix $\boldsymbol{A} = \mathbf{R}^{M\times N}$, $m \in [M]$, $n \in [N]$, regularization parameters γ, τ, and stopping tolerance ϵ

Initialize $\mu = 1/\tau$, $\boldsymbol{X}^0 = \boldsymbol{W}^0 = \boldsymbol{Y}^0 = (mn/MN)\boldsymbol{e}\boldsymbol{e}^T$

for $l = 0, 1, \ldots$ until converged **do**

Step 1. Update $\boldsymbol{Q}^{l+1}$: $\boldsymbol{Q}^{l+1} = P_\Omega\left(\boldsymbol{X}^l - \boldsymbol{Y}^l + \mu\boldsymbol{\Lambda}^l_{\boldsymbol{Q}}\right)$

Step 2. Update $\boldsymbol{X}^{l+1}$: Take SVD of

$$\tilde{\boldsymbol{X}}^l = 1/3\left(\boldsymbol{Y}^l + \boldsymbol{Q}^{l+1} + \boldsymbol{Z}^l + \boldsymbol{W}^l - \mu(\boldsymbol{\Lambda}^l_{\boldsymbol{Q}} + \boldsymbol{\Lambda}^l_{\boldsymbol{Z}} + \boldsymbol{\Lambda}^l_{\boldsymbol{W}})\right) = \boldsymbol{U}(\text{Diag}\boldsymbol{x})\boldsymbol{V}^T.$$

Apply soft thresholding: $\boldsymbol{X}^{l+1} = \boldsymbol{U}(\text{Diag}S_{\tau/3}(\boldsymbol{x}))\boldsymbol{V}^T$.

Step 3. Update $\boldsymbol{Y}^{l+1}$: $y^{l+1}_{ij} = \max\left\{\left[\boldsymbol{X}^{l+1} - \boldsymbol{Q}^{l+1} - \gamma\boldsymbol{e}\boldsymbol{e}^T\mu + \boldsymbol{\Lambda}^l_{\boldsymbol{Q}}\mu\right]_{ij}, 0\right\}$ for all $i \in [M]$, $j \in [N]$.

Step 4. Update $\boldsymbol{W}^{l+1}$: $\boldsymbol{W}^{l+1} = \tilde{\boldsymbol{W}}^l + \alpha^l\boldsymbol{e}\boldsymbol{e}^T$, where $\tilde{\boldsymbol{W}}^l = \boldsymbol{X}^{l+1} + \mu\boldsymbol{\Lambda}^l_{\boldsymbol{W}}$ and $\alpha^l = (mn - \boldsymbol{e}^T\tilde{\boldsymbol{W}}^l\boldsymbol{e})/(MN)$.

Step 5. Update variable $\boldsymbol{Z}^{l+1}$: $z^{l+1}_{ij} = \min\{\max\left\{\left[\boldsymbol{X}^{l+1} + \mu\boldsymbol{\Lambda}^l_{\boldsymbol{Z}}\right]_{ij}, 0\right\}, 1\}$ for all $i \in [M]$, $j \in [N]$.

Step 6. Update dual variables:

$$\begin{aligned}\boldsymbol{\Lambda}^{l+1}_{\boldsymbol{Q}} &= \boldsymbol{\Lambda}^l_{\boldsymbol{Q}} + \tau(\boldsymbol{X}^{l+1} - \boldsymbol{Y}^{l+1} - \boldsymbol{Q}^{l+1}),\\ \boldsymbol{\Lambda}^{l+1}_{\boldsymbol{W}} &= \boldsymbol{\Lambda}^l_{\boldsymbol{W}} + \tau(\boldsymbol{X}^{l+1} - \boldsymbol{W}^{l+1}),\\ \boldsymbol{\Lambda}^{l+1}_{\boldsymbol{Z}} &= \boldsymbol{\Lambda}^l_{\boldsymbol{Z}} + \tau(\boldsymbol{X}^{l+1} - \boldsymbol{Z}^{l+1}).\end{aligned}$$

Step 7. Calculate primal and dual residuals

$$\begin{aligned}r_p &= \max\{\|\boldsymbol{X}^{l+1} - \boldsymbol{Z}^{l+1}\|_F, \|\boldsymbol{X}^{l+1} - \boldsymbol{W}^{l+1}\|_F, \|\boldsymbol{X}^{l+1} - \boldsymbol{Y}^{l+1} - \boldsymbol{Q}^{l+1}\|_F\}/\|\boldsymbol{X}^{l+1}\|_F,\\ r_d &= \max\{\|\boldsymbol{Z}^{l+1} - \boldsymbol{Z}^{l}\|_F, \|\boldsymbol{W}^{l+1} - \boldsymbol{W}^{l}\|_F, \|\boldsymbol{Q}^{l+1} - \boldsymbol{Q}^{l}\|_F\}/\|\boldsymbol{X}^{l+1}\|_F.\end{aligned}$$

if $\max(r_p, r_d) < \epsilon$ **then** algorithm converged.

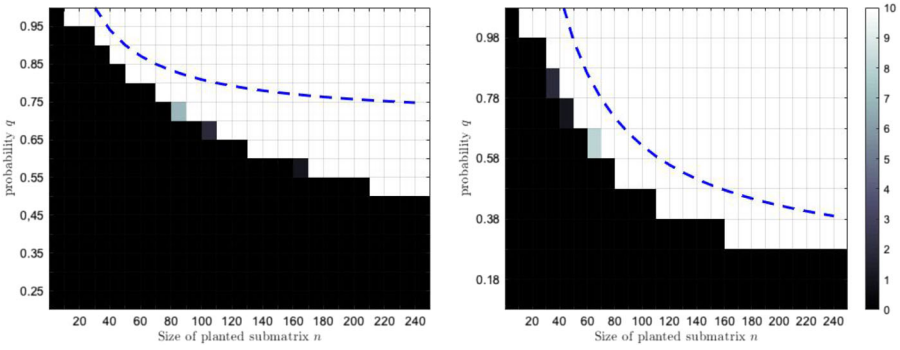
end if

end for

[1] A MATLAB implementation of Algorithm 1 is available from http://bpames.people.ua.edu/software and an R implementation of Algorithm 1 is available from the Comprehensive R Archive Network (CRAN) as the package admmDensestSubmatrix.

5.2 Random Matrices

We empirically verified the theoretical phase transitions provided by Theorem 3.1 using matrices randomly sampled from the planted dense subgraph model with fixed noise edge probability p and varied the submatrix size n and in-submatrix probability q. We perform two sets of experiments: one where the matrix is sparse outside the planted submatrix and another when the noise obscuring the planted submatrix is relatively dense. For the dense graph simulations, we choose $p = 0.25$ and $q \in \{0.25, 0.30, \ldots, 0.95, 1\}$. In the sparse experiments, we choose $p = 1/\sqrt{N}$ and $q = tp$ for ten equally spaced t spanning the interval $[2, \sqrt{N}]$. For each set of simulations, we vary $n \in \{10, 20, 30, \ldots, 240, 250\}$ and set $m = 2n$. In the sparse experiments, we have $M = N = 1000$ and we use $M = N = 500$ in dense experiments. In both the dense and sparse graph simulations, we generate 10 matrices according to the planted dense submatrix model for each choice of the parameters q and n (with remaining parameters p, M, and N chosen as described above). We call Algorithm 1 to solve the instance of Eq. 6 corresponding to each randomly sampled matrix. The regularization parameter $\gamma = 6/(q - p)n$, augmented Lagrangian parameter $\tau = 0.35$, and stopping tolerance $\epsilon = 10^{-4}$ are used in each call of Algorithm 1. We declared the planted submatrix to be recovered if the relative error $\|\boldsymbol{X}^* - \boldsymbol{X}^0\|_F / \|\boldsymbol{X}^0\|_F$ is less than 10^{-3}, where $\boldsymbol{X}^*$ is the solution returned by Algorithm 1 and $\boldsymbol{X}^0$ is the matrix representation of the planted submatrix. The empirical probability of recovery of planted submatrices is plotted in Fig. 1. The color of a square indicates rate of recovery in the corresponding simulations, with black corresponding to 0 and white corresponding to 10 recoveries out of 10 trials. The dashed curves show the phase transition to perfect recovery predicted by Theorem 3.1. The empirical recovery rates observed in these trials closely match that predicted by Theorem 3.1. The discrepancy between the observed phase transition and that more conservatively predicted by Theorem 3.1 is due to the presence of the logarithmic terms in Eq. 3; a slight modification of our proof to follow that of [31, Theorem 7]



(a) Dense case: $A = 500 \times 500$, $p = 0.25$ (b) Sparse case: $A = 1000 \times 1000$, $p = 1/\sqrt{N}$

Fig. 1 Recovery rates for randomly generated matrices (a) Dense 500×500 matrices with probability of adding nonzero entry outside of planted submatrix equal to $p = 0.25$. (b) Sparse 1000×1000 matrices with probability of adding nonzero entry outside of planted submatrix equal to $p = 1/\sqrt{1000}$

eliminates these terms when p and q are constants and the gap $q - p$ is sufficiently large.

5.2.1 Collaboration and Communication Networks

We also applied our algorithm to identify communities in networks taken from the 10th DIMACs Implementation Challenge, which focused on graph partitioning and clustering [54, 55]. The first graph (JAZZ) represents a collaboration network with 198 musicians and 2742 edges, and was compiled by [56]. Here, two musicians are connected if they have performed together. Earlier studies [10] showed that this network contains a cluster of 100 musicians. We apply Algorithm 1 to the adjacency matrix of this network with regularization parameter $\tau = 0.85$, stopping tolerance $\epsilon = 10^{-2}$, and $m = n = 100$. Our algorithm converges to the dense submatrix representing this community after 50 iterations. Figure 2 is a visualization of this network using the software package Gephi [57] and the ForceAtlas2 algorithm [58]. The `statistics` function of Gephi is used to identify three communities within this network, including the community of size $n = 100$ identified by Algorithm 1.

We also consider the graph (EMAIL) representing the network of e-mail interchanges between faculty, researchers, technicians, managers, administrators, and graduate students of the Univeristy Rovira i Virgili (Tarragona). Two individuals are connected if they exchanged an e-mail. There are 1133 nodes and 5451 edges. From [59], we know that the EMAIL graph has a dense subgraph of 289 vertices, representing a community of 289; additionally, we can identify 7 clusters using the `statistics` function of Gephi corresponding to academic units within this university, including this community. Applying Algorithm 1 with $m = n = 289$, $\tau = 0.35$, and stopping criteria $\epsilon = 10^{-2}$ finds this subgraph in 15 iterations. The results of these analyses are summarized in Table 1.

(a) JAZZ Network (b) Extracted communities

Fig. 2 JAZZ Network. Each color corresponds to membership in one of 3 clusters, isolated in the plot on the right

Table 1 Densest subgraphs extracted with ADMM

Graph	Number of vertices	Number of edges	Size of dense subgraph	Running time
JAZZ	198	2742	100	0.605735 s
EMAIL	1133	5451	289	20.139186 s

6 Conclusions

We have presented an analysis of new convex relaxations for the densest subgraph and submatrix problems and have established sufficient conditions under which the optimal solution of the original combinatorial problem coincides with that of these relaxations. In particular, these sufficient conditions characterize a signal-to-noise ratio (SNR) for matrices sampled from a particular distribution of random matrices, such that if this ratio is sufficiently large then we can expect to recover the combinatorial solution from the solution of the relaxation. Here, we expect perfect recovery if the strength of signal, as measured by the gap between probabilities of existence of nonzero within-group edges and out-group entries, is sufficiently larger than noise, as measured by variability of presence of nonzero entries. Further, the SNR corresponding to this phase transition to perfect recovery matches the current state of the art identified in the previous literature (up to constant and logarithmic terms); see [43].

This recovery guarantee provides a *sufficient* condition for perfect recovery of the planted subgraph or submatrix. It would be very interesting to determine if this condition is also *necessary*. For example, we establish that we have perfect recovery of planted dense k-submatrices and subgraphs if $k \geq \Omega(N)$ when the probabilities q, p are sufficiently large constants. It is unclear if it is possible, either using our relaxation or some other method, to efficiently recover planted submatrices and subgraphs of size $O(n^{1/2-\epsilon})$ for some $\epsilon > 0$.

A secondary open problem focuses on efficient solution of the proposed convex relaxations. We currently solve these problems using a multi-block variant of the ADMM. Each iteration of this algorithm requires $O(N^3)$ arithmetic operations; the bulk of these operations are used by the calculation of the singular value decomposition used to update $\boldsymbol{X}$. This per-iteration cost scales unfavorably when N is large. The recent manuscript by Sotirov [60] proposed a coordinate descent heuristic for the densest k-subgraph, with empirical evidence that this heuristic efficiently solves large-scale instances of the densest k-subgraph problem. However, sufficient conditions for perfect recovery of a planted dense submatrix have not yet been established for this method. Further research is needed to design efficient and scalable algorithms, i.e., with per-iteration cost $O(N)$, with provable theoretical guarantees of recovery for the solution of the densest subgraph and submatrix problems.

Funding Information B. Ames was supported by NSF DMS-2012554, University of Alabama CyberSeed Grant SP14572, and University of Alabama Research Grants RG14678 and RG14838.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Appendix 1. Proof of Lemma 4.3

The proof is virtually identical to that of [32, Lemma 4.5]. We decompose $\boldsymbol{Q}$ as $\boldsymbol{Q} = \lambda \boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{H} + \boldsymbol{y}\boldsymbol{e}^T + \boldsymbol{e}\boldsymbol{z}^T$, where $\boldsymbol{H}$ is matrix defined by $H_{ij} = 1$ if $ij \in \Omega$ and $H_{ij} = 0$ otherwise. We can further decompose $\boldsymbol{Q}$ as $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2 + \boldsymbol{Q}_3 + \boldsymbol{Q}_4$, where $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$, $\boldsymbol{Q}_3$, $\boldsymbol{Q}_4$ are constructed as below.

We first bound $\boldsymbol{Q}_1 := (1-q)\boldsymbol{e}\boldsymbol{e}^T - \boldsymbol{H}$. Note that $\boldsymbol{Q}_1$ has i.i.d. mean zero entries, with variance $\sigma^2 = \sigma_q^2$ and values either $1-q$ with probability q or $-q$ with probability $1-q$. Applying Eq. 23 with $B = 1$ and $t = N$,

$$\|\boldsymbol{Q}_1\| = O\left(\max\left\{\sqrt{\sigma_q^2 n}, \sqrt{\log N}\right\}\right) \tag{27}$$

w.h.p. Next, we let $\boldsymbol{Q}_2 := \frac{1}{n}(\bar{\boldsymbol{\mu}}\boldsymbol{e}^T - (1-q)n\boldsymbol{e}\boldsymbol{e}^T)$. Note that

$$\|\boldsymbol{Q}_2\|_2 = \tfrac{1}{n}\|(\bar{\boldsymbol{\mu}} - (1-q)n\boldsymbol{e})\boldsymbol{e}^T\|_2 \le \tfrac{1}{n}\|\bar{\boldsymbol{\mu}} - (1-q)n\boldsymbol{e}\|\|\boldsymbol{e}\| = \tfrac{1}{\sqrt{n}}\|\bar{\boldsymbol{\mu}} - (1-q)n\boldsymbol{e}\|.$$

Applying Eq. 14 shows that $\bar{\mu}_i - (1-q)n \le 6\max\left\{\sqrt{\sigma_q^2 n \log N}, \log N\right\}$ for all $i \in \bar{U}$ w.h.p. It follows that

$$\|\boldsymbol{Q}_2\| \le 6\sqrt{\frac{m}{n}}\max\left\{\sqrt{\sigma_q^2 n \log N}, \log N\right\} \tag{28}$$

w.h.p. Next, let $Q_3 := \frac{1}{m}\boldsymbol{e}\bar{\boldsymbol{v}}^T - (1-q)\boldsymbol{e}\boldsymbol{e}^T$. An identical argument shows that

$$\|\boldsymbol{Q}_3\| \le 6\sqrt{\frac{n}{m}}\max\left\{\sqrt{\sigma_q^2 m \log N}, \log N\right\} \tag{29}$$

w.h.p. Finally, we let

$$\boldsymbol{Q}_4 := \left(\frac{(1-q)mn - \bar{\boldsymbol{\mu}}^T\boldsymbol{e}}{mn}\right)\boldsymbol{e}\boldsymbol{e}^T.$$

It is easy to confirm that $\gamma(\boldsymbol{Q}_1 + \boldsymbol{Q}_2 + \boldsymbol{Q}_3 + \boldsymbol{Q}_4) = \boldsymbol{W}(\bar{U}, \bar{V})$. Applying Eq. 16 shows that

$$\|\boldsymbol{Q}_4\| \le \frac{1}{\sqrt{mn}}6\max\left\{\sqrt{\sigma_q^2 mn \log N}, \log N\right\} \tag{30}$$

w.h.p. Combining Eqs. 27, 28, 29, and 30 establishes that

$$\|\boldsymbol{Q}\| \le \sum_{i=1}^{4}\|\boldsymbol{Q}_i\| = O\left(\max\left\{\sqrt{\sigma_q^2 n \log N}, \sqrt{\frac{n}{m}}\log N\right\}\right)$$

w.h.p., as required.

Appendix 2. Proof of Lemma 4.4

To obtain the desired bound on $\boldsymbol{S}$, we first approximate $\boldsymbol{S}$ with a random matrix with mean zero entries. In particular, we let $\tilde{\boldsymbol{S}}_1$ be the random matrix constructed as follows. For all $(i, j) \notin \bar{U} \times \bar{V}$ such that $ij \notin \Omega$, or $(i, j) \in ([M] - \bar{U}) \times ([N] - \bar{V})$ such that $(i, j) \in \Omega$, we let $[\tilde{\boldsymbol{S}}_1]_{ij} = S_{ij}$. All remaining entries of $\tilde{\boldsymbol{S}}_1$ are sampled independently from the generalized Bernoulli distribution $\mathcal{B}$, where x sampled from $\mathcal{B}$ satisfy

$$x = \left\{ \begin{array}{c} \lambda, \text{ with probability } p, \\ -\lambda\, \tilde{\sigma}_p^2, \text{ with probability } 1 - p. \end{array} \right\}$$

Note that $\tilde{\boldsymbol{S}}_1$ is a random matrix with i.i.d. mean zero entries sampled independently from $\mathcal{B}$ by our choice of $\boldsymbol{W}$. Applying Lemma 4.2 shows that

$$\|\tilde{\boldsymbol{S}}_1\| = O\left(\max\left\{\sqrt{\tilde{\sigma}_p^2 N}, \sqrt{B \log N}\right\}\right) \tag{31}$$

w.h.p. The remainder of the proof establishes that $\boldsymbol{S}$ is well-approximated by $\tilde{\boldsymbol{S}}_1$, i.e., we complete the proof by bounding the norm of the error $\boldsymbol{S} - \tilde{\boldsymbol{S}}_1$. We begin with the error in the $\bar{U} \times \bar{V}$ block. Let $\tilde{\boldsymbol{S}}_2 = -\tilde{\boldsymbol{S}}_1(\bar{U}, \bar{V})$. Applying Lemma 4.2 with $t = N$ again shows that

$$\|[\boldsymbol{S} - \tilde{\boldsymbol{S}}_1](\bar{U}, \bar{V})\| = \|\tilde{\boldsymbol{S}}_2\| = O\left(\max\left\{\sqrt{\tilde{\sigma}_p^2 n}, \sqrt{B \log N}\right\}\right) \tag{32}$$

w.h.p. We define $\tilde{\boldsymbol{S}}_3$ by

$$[\tilde{\boldsymbol{S}}_3]_{ij} = \left\{ \begin{array}{c} -\frac{v_j}{m - v_j} + \frac{p}{1-p}, \text{ if } (i, j) \in \Omega,\ i \in \bar{U}, j \in [N] - \bar{V}, \\ 0, \text{ otherwise.} \end{array} \right\}$$

To bound the norm of $\tilde{\boldsymbol{S}}_3$, we will use the following lemma, which provides a bound on the spectral norm of random matrices of this form.

Lemma B.1 *Let $\boldsymbol{A}$ be an $n \times N$ matrix whose entries are chosen according to $\mathcal{B}$ with $n \leq N$. Let $\tilde{\boldsymbol{A}}$ be the random matrix defined by*

$$[\tilde{A}]_{ij} := \begin{cases} 1, & \text{if } A_{ij} = 1, \\ \frac{-n_j}{n - n_j}, & \text{if } A_{ij} = -\tilde{\sigma}_p^2, \end{cases}$$

where n_j is the number of 1's in the jth column of $\boldsymbol{A}$. Then, there are constants $c_1, c_2 > 0$ such that

$$\Pr\left(\|\boldsymbol{A} - \tilde{\boldsymbol{A}}\| \geq c_1 \max\left\{\sqrt{\tilde{\sigma}_p^2 N} \log N, (\log N)^{3/2}\right\}\right) \leq c_2 N^{-5}.$$

The proof of Lemma B.1 follows a similar argument to that of [31, Theorem 4] and is included as Appendix C. It is easy to see that the nonzero block $\tilde{\boldsymbol{S}}_3$ has form $\boldsymbol{A} - \tilde{\boldsymbol{A}}$ as in the hypothesis of Lemma B.1. It follows that

$$\|[\boldsymbol{S} - \boldsymbol{S_1}](\bar{U}, [N] - \bar{V})\| = \|\tilde{\boldsymbol{S}}_3\| = O\left(\max\left\{\sqrt{\tilde{\sigma}_p^2 N} \log N, (\log N)^{3/2}\right\}\right) \tag{33}$$

w.h.p. Similarly, we define the final correction matrix by

$$[\tilde{S}_4]_{ij} = \left\{ \begin{array}{c} -\frac{\mu_j}{n-\mu_j} + \frac{p}{1-p}, \text{ if } (i,j) \in \Omega,\ i \in [M] - \bar{U},\ j \in \bar{V}, \\ 0, \text{ otherwise.} \end{array} \right\}$$

Applying Lemma B.1 to the transpose of $\tilde{S}_4$, we have

$$\|\tilde{S}_4\| = O\left(\max\left\{\sqrt{\tilde{\sigma}_p^2 M}\log M, (\log M)^{3/2}\right\}\right) = O\left(\max\left\{\sqrt{\tilde{\sigma}_p^2 N}\log N, (\log N)^{3/2}\right\}\right) \quad (34)$$

w.h.p. Combining Eqs. 31, 32, 33, and 34 and applying the union bound completes the proof.

Appendix 3. Proof of Lemma B.1

The result relies on an application of the Matrix Bernstein Inequality; see [61, Theorem 6.1.1] and [62, Theorem 1.6] for further details. We first state the necessary bound on the spectral norm of the sum of finitely many independent, bounded random matrices.

Theorem C.1 (**Matrix Bernstein Inequality**) *Let $\{\boldsymbol{S}_k\}$ be a finite sequence of independent $d_1 \times d_2$ random matrices such that $\mathbb{E}[\boldsymbol{S}_k] = \boldsymbol{0}$ and $\|\boldsymbol{S}_k\| \le L$ for all k almost surely. Let $\boldsymbol{Z} := \sum_k \boldsymbol{S}_k$ and let $v(\boldsymbol{Z})$ denote the matrix variance defined by*

$$v(\boldsymbol{Z}) = \max\left\{\|\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^*]\|, \|\mathbb{E}[\boldsymbol{Z}^*\boldsymbol{Z}]\|\right\} = \max\left\{\left\|\sum_k \mathbb{E}[\boldsymbol{S}_k\boldsymbol{S}_k^*]\right\|, \left\|\sum_k \mathbb{E}[\boldsymbol{S}_k^*\boldsymbol{S}_k]\right\|\right\}. \quad (35)$$

Then,

$$Pr(\|\boldsymbol{Z}\| \ge t) \le (d_1 + d_2)\exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3}\right) \quad (36)$$

for all $t > 0$.

The remainder of the proof consists of a specialization of this inequality to the special case $\boldsymbol{Z} = \boldsymbol{A} - \tilde{\boldsymbol{A}}$. Indeed, let

$$\boldsymbol{S}_j = \boldsymbol{d}_j \boldsymbol{e}_j^T, \quad (37)$$

where $\boldsymbol{d}_j := [\boldsymbol{A} - \tilde{\boldsymbol{A}}]_j$ denotes the jth column of $\boldsymbol{A} - \tilde{\boldsymbol{A}}$ and $\boldsymbol{e}_j$ denotes the jth standard basis vector. It is clear that $\boldsymbol{Z} = \boldsymbol{A} - \tilde{\boldsymbol{A}} = \sum_{j=1}^N \boldsymbol{S}_j$. It remains to estimate an upper bound L on the spectral norms of the matrices $\{\boldsymbol{S}_j\}$ and an upper bound on the variance $v(\boldsymbol{Z})$. Once we have estimated these quantities, we will substitute them into Eq. 36 to complete the proof.

We begin with the following estimate on L, which is an immediate consequence of the standard Bernstein Inequality (Eq. 13).

Lemma C.1 *There is a constant $c_1 > 0$ such that matrices $\{\boldsymbol{S}_j\}$ defined by* Eq. 37 *satisfy*

$$\|\boldsymbol{S}_j\| \le L := c_1\sqrt{\max\left\{1, \tilde{\sigma}_p^2\right\}\log N}$$

for all $j = 1, 2, \ldots, N$ *with probability at least* $1 - 2N^{-5}$.

Proof Fix $j \in \{1, \ldots, N\}$. Note that the $\|\boldsymbol{S}_j\| = \|\boldsymbol{d}_j \boldsymbol{e}_j^T\| = \|\boldsymbol{d}_j\|\|\boldsymbol{e}_j\| = \|\boldsymbol{d}_j\|$. Moreover, the Bernstein Inequality (Eq. 13) implies that

$$\|\boldsymbol{d}_j\|^2 = \frac{(n_j - pn)^2}{(1-p)^2(n-n_j)} \leq \frac{36\max\{p(1-p)n\log N, \log^2 N\}}{(1-p)^2(n - 6\max\{\sqrt{p(1-p)n\log N}, \log N\})} = O\Big(\max\{1, \tilde{\sigma}_p^2\}\log N\Big)$$

with probability at least $1 - 2N^{-6}$, where the last inequality follows from the fact that $n - n_j = O(n)$ w.h.p. (by Eq. 13) and $\log N = O(n)$ (by the gap assumption). Taking the square root completes the proof. □

We next bound the matrix variance $v(\boldsymbol{Z})$.

Lemma C.2 *The matrix* $\boldsymbol{Z} = \sum_j \boldsymbol{S}_j$ *defined by* Eq. 37 *satisfies* $v(\boldsymbol{Z}) \leq c\tilde{\sigma}_p^2 N$ *for any constant* $c > 0$ *satisfying* $n - n_j > (1/c)n$ *for all* j.

Proof It suffices to construct upper bounds on each of $\|\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]\|$ and $\|\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}]\|$. We begin with the latter. Note that $\boldsymbol{S}_j^T\boldsymbol{S}_j^T = (\boldsymbol{d}_j^T\boldsymbol{d}_j)\boldsymbol{e}_j\boldsymbol{e}_j^T$. This implies that $\boldsymbol{Z}^T\boldsymbol{Z}$ is a diagonal matrix with jth diagonal entry equal to $\|\boldsymbol{d}_j\|^2$. It follows that $\|\mathbb{E}(\boldsymbol{Z}^T\boldsymbol{Z})\| = \max_j \mathbb{E}[\|\boldsymbol{d}_j\|^2]$. For each $j = 1, 2, \ldots, N$, we have

$$\mathbb{E}[\|\boldsymbol{d}_j\|^2] = \mathbb{E}\left[\frac{(n_j - pn)^2}{(1-p)^2(n-n_j)}\right] \leq \frac{c}{(1-p)^2 n}\mathbb{E}[(n_j - pn)^2] = \frac{cp(1-p)n}{(1-p)^2 n} = c\tilde{\sigma}_p^2, \quad (38)$$

where the inequality follows from the assumption that $n - n_j \geq (1/c)n$ and the second to last inequality follows from the fact that $\mathbb{E}[(n_j - pn)^2]$ is equal to the variance of the binomial variable n_j. This implies that

$$\|\mathbb{E}[\boldsymbol{Z}^T\boldsymbol{Z}]\| \leq c\tilde{\sigma}_p^2. \quad (39)$$

On the other hand, $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] = \sum_{j=1}^N \mathbb{E}[\boldsymbol{S}_j\boldsymbol{S}_j^T] = \sum_{j=1}^N \mathbb{E}[\boldsymbol{d}_j\boldsymbol{d}_j^T]$. It follows immediately that

$$\|\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]\| \leq \sum_{j=1}^N \|\mathbb{E}[\boldsymbol{d}_j\boldsymbol{d}_j^T]\| \leq \sum_{j=1}^N \mathbb{E}[\|\boldsymbol{d}_j\boldsymbol{d}_j^T\|] = \sum_{j=1}^N \mathbb{E}[\|\boldsymbol{d}_j\|^2] \quad (40)$$

by the triangle inequality and Jensen's inequality. Applying Eq. 38, we see that

$$\|\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]\| \leq cN\tilde{\sigma}_p^2. \quad (41)$$

Substituting Eq. 39 and 41 into the formula for $v(\boldsymbol{Z})$, we see that we have $v(\boldsymbol{Z}) \leq cN\tilde{\sigma}_p^2$. □

We are now ready to complete the proof of Lemma B.1. Let's consider the following cases. First, suppose that $\tilde{\sigma}_p^2 N \geq \log N$ and let $t = \tilde{c}\sqrt{\tilde{\sigma}_p^2 N \log N}$ in Eq. 36. Recall that applying the Bernstein Inequality (Eq. 13) to each binomial variable n_j implies that there is a constant c such that $n - n_j \geq (1/c)n$ for all j with probability

at least $1-2N^{-5}$. This implies that we have $v(\boldsymbol{Z}) \leq c\tilde{\sigma}_p^2 N$ with the same probability. Substituting into Eq. 36, along with the choice of L from Lemma C.1, we see that

$$Pr(\|\boldsymbol{Z}\| \geq t) \leq (N+n)\exp\left(-\frac{\tilde{c}^2\tilde{\sigma}_p^2 N\log^2 N/2}{c\tilde{\sigma}_p^2 N+\tilde{c}\sqrt{B}\tilde{\sigma}_p^2 N\log N/3}\right)$$

using the assumption that $\sqrt{\log N} \leq \sqrt{\tilde{\sigma}_p^2 N}$. Rearranging further, we see that

$$Pr(\|\boldsymbol{Z}\| \geq t) \leq (N+n)\exp\left(-\frac{\tilde{c}^2}{c+\tilde{c}\sqrt{B}}\log N\right) \leq (N+n)\exp(-7\log N) \leq 2N^{-6},$$

if we choose $\tilde{c}$ large enough that $\tilde{c}^2/(c+\tilde{c}\sqrt{B}) > 7$ (which is possible if we impose the assumption that $B = O(1)$).

Next, consider the case that $\log N > \tilde{\sigma}_p^2 N$ and let $t = \tilde{c}\log^{3/2}(N)$. Then, the Matrix Bernstein Inequality (Eq. 36) implies that

$$Pr(\|\boldsymbol{Z}\| \geq t) \leq (N+n)\exp\left(\frac{-\tilde{c}\log^3 N/2}{c\tilde{\sigma}_p^2 N+\tilde{c}\sqrt{B}\log^2 N/3}\right) \leq (N+n)\exp(-7\log N) \leq 2N^{-6}$$

for any $\tilde{c}$ satisfying $\tilde{c}^2/(c+\tilde{c}\sqrt{B}) > 7$. Combining the two cases, we see that there are constants $c_1, c_2 > 0$ such that

$$\|\boldsymbol{Z}\| = \|\boldsymbol{A}-\tilde{\boldsymbol{A}}\| \leq c_1 \max\left\{\sqrt{\tilde{\sigma}_p^2 N\log N}, \log N\right\}\sqrt{\log N}$$

with probability at least $1-c_2N^{-5}$. This completes the proof.

References

1. Karp R (1972) Reducibility among combinatorial problems. Complexity of Computer Computations 40(4):85–103
2. Alon N, Arora S, Manokaran R, Moshkovitz D, Weinstein O (2011) Inapproximability of densest κ-subgraph from average case hardness. Unpublished manuscript 1
3. Feige U (2002) Relations between average case complexity and approximation complexity. In: Proceedings of the thiry-fourth annual ACM symposium on theory of computing. ACM, pp 534–543
4. Khot S (2006) Ruling out PTAS for graph min-bisection, dense k-subgraph, and bipartite clique. SIAM J Comput 36(4):1025–1071
5. Henzinger MR, Motwani R, Silverstein C (2002) Challenges in web search engines. In: ACM SIGIR forum, vol 36. ACM, pp 11–22
6. Gibson D, Kumar R, Tomkins A (2005) Discovering large dense subgraphs in massive graphs. In: Böhm K, Jensen CS, Haas L M, Kersten ML, Larson P, Ooi BC (eds) Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005. ACM, pp 721–732
7. Angel A, Koudas N, Sarkas N, Srivastava D (2012) Dense subgraph maintenance under streaming edge weight updates for real-time story identification. Proc. VLDB Endow. 5(6):574–585
8. Gajewar A, Das Sarma A (2012) Multi-skill collaborative teams based on densest subgraphs. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM, pp 165–176
9. Tsourakakis C (2015) The k-clique densest subgraph problem. In: Proceedings of the 24th international conference on world wide web, pp 1122–1132. International World Wide Web Conferences Steering Committee
10. Tsourakakis C, Bonchi F, Gionis A, Gullo F, Tsiarli M (2013) Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 104–112

11. Abbe E, Bandeira A, Hall G (2016) Exact recovery in the stochastic block model. IEEE Trans Inf Theory 62(1):471–487
12. Ailon N, Chen Y, Xu H (2013) Breaking the small cluster barrier of graph clustering. In: International conference on machine learning, pp 995–1003
13. Ames B, Vavasis S (2014) Convex optimization for the planted k-disjoint-clique problem. Math Program 143(1-2):299–337
14. Ames B (2014) Guaranteed clustering and biclustering via semidefinite programming. Math Program 147(1-2):429–465
15. Amini A, Levina E (2018) On semidefinite relaxations for the block model. Ann Stat 46(1):149–179
16. Cai T, Li X (2015) Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. Ann Stat 43(3):1027–1059
17. Chen Y, Jalali A, Sanghavi S, Xu H (2014) Clustering partially observed graphs via convex optimization. The Journal of Machine Learning Research 15(1):2213–2238
18. Chen Y, Sanghavi S, Xu H (2014) Improved graph clustering. IEEE Trans Inf Theory 60(10):6440–6455
19. Chen Y, Xu J (2014) Statistical-computational phase transitions in planted models: the high-dimensional setting. In: International conference on machine learning, pp 244–252
20. Guédon O, Vershynin R (2015) Community detection in sparse networks via Grothendieck's inequality. Probab Theory Relat Fields, pp 1–25
21. Hajek B, Wu Y, Xu J (2015) Achieving exact cluster recovery threshold via semidefinite programming. In: IEEE international symposium on information theory. IEEE, pp 1442–1446
22. Lei J, Rinaldo A (2015) Consistency of spectral clustering in stochastic block models. Ann Stat 43(1):215–237
23. Mathieu C, Schudy W (2010) Correlation clustering with noisy input. In: ACM-SIAM symposium on discrete algorithms, pp 712–728. Society for Industrial and Applied Mathematics
24. Nellore A, Ward R (2015) Recovery guarantees for exemplar-based clustering. Inf Comput 245:165–180
25. Oymak S, Hassibi B (2011) Finding dense clusters via "low rank + sparse" decomposition. arXiv:1104.5186
26. Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. Ann Stat 39(4):1878–1915
27. Qin T, Rohe K (2013) Regularized spectral clustering under the degree-corrected stochastic blockmodel. In: Advances in neural information processing systems, pp 3120–3128
28. Vinayak R, Oymak S, Hassibi B (2014) Sharp performance bounds for graph clustering via convex optimization. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 8297–8301
29. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Social networks 5(2):109–137
30. Li X, Chen Y, Xu J (2018) Convex relaxation methods for community detection. arXiv:1810.00315
31. Ames B, Vavasis SA (2011) Nuclear norm minimization for the planted clique and biclique problems. Mathematical Programming 129(1):69–89
32. Ames B (2015) Guaranteed recovery of planted cliques and dense subgraphs by convex relaxation. J Optim Theory Appl 167(2):653–675
33. Bomze IM, Budinich M, Pardalos PM, Pelillo M (1999) The maximum clique problem. In: Handbook of combinatorial optimization. Springer, pp 1–74
34. Pardalos PM, Xue J (1994) The maximum clique problem. Journal of global Optimization 4(3):301–328
35. Deshpande Y, Montanari A (2015) Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. Found Comput Math 15(4):1069–1128
36. Montanari A (2015) Finding one community in a sparse graph. J Stat Phys 161(2):273–299
37. Hajek B, Wu Y, Xu J (2017) Information limits for recovering a hidden community. IEEE Trans Inf Theory 63(8):4729–4745
38. Hajek B, Wu Y, Xu J (2016) Semidefinite programs for exact recovery of a hidden community. In: Conference on learning theory, pp 1051–1095
39. Saad H, Nosratinia A (2018) Belief propagation with side information for recovering a single community. In: 2018 IEEE international symposium on information theory (ISIT). IEEE, pp 1271–1275

40. Hajek B, Wu Y, Xu J (2018) Recovering a hidden community beyond the Kesten–Stigum threshold in $O(|E|log^*|V|)$ time. J Appl Probab 55(2):325–352. https://doi.org/10.1017/jpr.2018.22. ISSN:0021-9002, 62H12 (60C05) 3832891 https://doi.org/10.1017/jpr.2018.22
41. Hajek B, Wu Y, Xu J (2017) Submatrix localization via message passing. J Machine Learn Res 18(1):6817–6868
42. Banks J, Moore C, Vershynin R, Verzelen N, Xu J (2018) Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. IEEE Trans Inform Theory
43. Brennan M, Bresler G, Huleihel W (2019) Universality of computational lower bounds for submatrix detection. arXiv:1902.06916
44. Chen Y, Sanghavi S, Xu H (2012) Clustering sparse graphs. In: Advances in neural information processing systems, pp 2204–2212
45. Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev 52(3):471–501
46. Alon N, Krivelevich M, Sudakov B (1998) Finding a large hidden clique in a random graph. Random Structures and Algorithms 13(3-4):457–466
47. Feige U, Ron D (2010) Finding hidden cliques in linear time. In: 21st international meeting on probabilistic, combinatorial, and asymptotic methods in the analysis of algorithms (AofA'10), pp 189–204. Discrete Mathematics and Theoretical Computer Science
48. Dekel Y, Gurel-Gurevich O, Peres Y (2014) Finding hidden cliques in linear time with high probability. Comb Probab Comput 23(1):29–49
49. Juels A, Peinado M (2000) Hiding cliques for cryptographic security. Des Codes Crypt 20(3):269–280
50. Golub G, Van Loan C (1996) Matrix computations. Johns Hopkins University Press
51. Boucheron S, Lugosi G, Massart P (2013) Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press
52. Bandeira AS, van Handel R (2014) Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Annals of Probability to appear
53. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J et al (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends®, in Machine learning 3(1):1–122
54. Bader DA, Meyerhenke H, Sanders P, Wagner D (2012) Graph partitioning and graph clustering. In: 10th DIMACS implementation challenge workshop
55. Bader DA, Meyerhenke H, Sanders P, Schulz C, Kappes A, Wagner D (2014) Benchmarking for graph clustering and partitioning. Encyclopedia of Social Network Analysis and Mining, pp 73–82
56. Gleiser P, Danoni L (2003) Community structure in jazz. Advances in Complex Systems 6(4):565–573
57. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: Third international AAAI conference on weblogs and social media
58. Jacomy M, Venturini T, Heymann S, Bastian M (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PloS one 9(6):e98,679
59. Tsourakakis C, Bonch F (2013) Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 104–112
60. Sotirov R (2019) On solving the densest k-subgraph problem on large graphs. arXiv:1901.06344
61. Tropp JA et al (2015) An introduction to matrix concentration inequalities. Foundations and Trends®, in Machine Learning 8(1-2):1–230
62. Tropp JA (2012) User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics 12(4):389–434