

Community detection with a subsampled semidefinite program

Journal Article**Author(s):**

Abdalla, Pedro; Bandeira, Afonso S.

Publication date:

2022-05-06

Permanent link:

<https://doi.org/10.3929/ethz-b-000547746>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Sampling Theory, Signal Processing, and Data Analysis 20, <https://doi.org/10.1007/s43670-022-00023-9>



Community detection with a subsampled semidefinite program

Pedro Abdalla¹ · Afonso S. Bandeira¹

Received: 24 September 2021 / Accepted: 12 April 2022
© The Author(s) 2022

Abstract

Semidefinite programming is an important tool to tackle several problems in data science and signal processing, including clustering and community detection. However, semidefinite programs are often slow in practice, so speed up techniques such as sketching are often considered. In the context of community detection in the stochastic block model, Mixon and Xie (IEEE Trans Inform Theory 67(10): 6832–6840, 2021) have recently proposed a sketching framework in which a semidefinite program is solved only on a subsampled subgraph of the network, giving rise to significant computational savings. In this short paper, we provide a positive answer to a conjecture of Mixon and Xie about the statistical limits of this technique for the stochastic block model with two balanced communities.

Keywords Community detection · Clustering algorithms · Graph and network theory · Semidefinite programming

Mathematics Subject Classification 94-XX

1 Introduction

Clustering problems are ubiquitous in data science. The main goal is to find a partition of the data into clusters in such form that the members in the same cluster are more

Communicated by Akram Aldroubi.

This article is part of the topical collection “Recent advances in computational harmonic analysis” edited by Dae Gwan Lee, Ron Levie, Johannes Maly and Hanna Veselovska.

✉ Pedro Abdalla
pedro.abdallateixeira@ifor.math.ethz.ch

Afonso S. Bandeira
bandeira@math.ethz.ch

¹ Department of Mathematics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Zürich, Switzerland

similar than the members in different clusters. At the same time it is necessary to balance the clusters sizes to avoid the trivial solution of one cluster per member.

A large body of work has focused on the stochastic block model, a random network model with a planted cluster structure, we refer the reader to [2] for a survey on recent developments. We will focus on case of two balanced communities. Let n be an even natural number and $G \sim \mathcal{G}(n; p, q)$ be a random graph on n nodes drawn as follows: Randomly partition the set of n vertices V in two equally sized communities $V = S_1 \cup S_2$. For every pair of vertices, an edge is placed with probability p if they belong to the same community S_i and with probability $q < p$ otherwise, all independent. The goal is to exactly recover the partition $\{S_1, S_2\}$ from the graph alone. Let the matrix $A \in \mathbb{R}^{n \times n}$ denote the adjacency matrix of the graph G . Considering a label vector $x \in \{\pm 1\}^n$ representing community membership of nodes¹. The maximum likelihood estimator for the node labels x is given by the program below [2],

$$\begin{aligned} \max_x \quad & x^T A x \\ \text{s.t.} \quad & \mathbf{1}^T x = 0 \\ & x \in \{\pm 1\}^n \end{aligned} \tag{1}$$

Here $\mathbf{1}$ denotes all-ones vector. Since it is well known that the problem (1) is NP-Hard [6], we consider the standard semidefinite relaxation [7].

$$\begin{aligned} \max_{X \in \mathbb{R}^{n \times n}} \quad & \text{Tr}(AX) \\ \text{s.t.} \quad & X_{ii} = 1 \\ & X \succeq 0 \\ & \text{Tr}(X\mathbf{J}) = 0 \end{aligned} \tag{2}$$

Where X is a surrogate variable for xx^T and \mathbf{J} denotes all-ones matrix. The following theorem gives the sharp phase transition for the community detection problem with two balanced communities.

Theorem 1 (Exact recovery threshold [1, 4, 8, 10]) *Let $G \sim \mathcal{G}(n; p, q)$ with $p = \alpha \frac{\log n}{n}$, $q = \beta \frac{\log n}{n}$ and planted communities $\{S_1, S_2\}$. Then,*

- (I) *For $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$, no algorithm can exactly recover the partition with high probability.*
- (II) *For $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, with high probability: The semidefinite program (2) has a unique solution given by $X^\natural = x^\natural(x^\natural)^T$ where x^\natural corresponds to the memberships of the true communities, thus achieving exact recovery.*

Although polynomial time, semidefinite programs tend to be computationally costly. A powerful tool to overcome computational complexity is that of sketching (we refer the reader to [13] for an instance of this idea in least squares, and [5, 14] for

¹ Note that there is a natural ambiguity in the labelling of each of the communities, thus the goal is best formulated in terms of recovering the partition; this corresponds of an ambiguity of global sign flip in x .

semidefinite optimization). In the particular framework addressed in this paper, Mixon and Xie [9] have recently proposed a sketching approach wherein a potentially significantly smaller semidefinite program is solved, its size depends on the community structure strength. Our main contribution is to resolve in the positive a conjecture in [9] regarding the dependency of the size of the resulting semidefinite program and the community structure strength. We now describe the sketching approach in [9], which consists of a three step process, and a tuning parameter $0 < \gamma < 1$.

- (Step 1) Given a graph with vertex set V . Subsample a smaller vertex set V^\natural by sampling each node in V independently at random with probability γ .
- (Step 2) Solve the community detection problem in the subgraph induced by V^\natural .
- (Step 3) For each node v not in V^\natural use a majority vote procedure among the neighbours of v in V^\natural to infer its community membership.

The main goal of this paper is to determine the minimum value of γ such that the approach above exactly recovers both communities with high probability. The computational savings come from the reduced size of the semidefinite program and so the parameter γ governs the computational cost of the algorithm (we refer the reader to [3] for the dependency of the computational cost of semidefinite programming on the number of variables).

Mixon and Xie [9] conjectured that, as long as

$$\gamma > \frac{2}{(\sqrt{\alpha} - \sqrt{\beta})^2},$$

the sketching approach works with high probability. Our main result provides a positive answer for this conjecture. In particular, for $\gamma = 1$, we recover the threshold in Theorem 1 (see part II).

2 An oracle bound

As described above, the sketching approach consists of three steps: Sampling, solving the community detection problem for a smaller sampled graph and then recovering the entire communities using a majority vote procedure. In this section, we analyze the Step 3 and prove that it works, for a certain range of the parameter γ , as long as we know the smaller communities in Step 2. The analysis is described in the proposition below, we refer to it as an oracle bound because it assumes the knowledge of the communities in Step 2.

Proposition 2 *Let $G \sim \mathcal{G}(n; p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$ satisfying $p > q$. Draw a vertex set V^\natural at random by sampling each node of the graph G independently at random with probability γ . Let R_1, R_2 be the planted communities in the sampled graph, i.e. $R_i = S_i \cap V^\natural$ for both $i \in \{1, 2\}$. Moreover, let $e(v, S)$ be the number of edges of G between the vertex v and the set*

$S \subset V(G)$ where $V(G)$ is the vertex set of the graph G . Now, consider

$$\begin{aligned} \hat{S}_1 &= R_1 \cup \{v \in V(G) \setminus V^\natural : e(v, R_1) > e(v, R_2)\}. \\ \hat{S}_2 &= R_2 \cup \{v \in V(G) \setminus V^\natural : e(v, R_2) > e(v, R_1)\}. \end{aligned}$$

Then there exists absolute constants $C, c > 0$ such that, with probability $1 - Cn^{-c((\alpha+\beta)^{\frac{\gamma}{2}} - \gamma\sqrt{\alpha\beta} - 1)}$, $(\hat{S}_1, \hat{S}_2) = (S_1, S_2)$. In particular, $(\hat{S}_1, \hat{S}_2) = (S_1, S_2)$ with probability $1 - o(1)$, as long as

$$\gamma > \frac{2}{(\sqrt{\alpha} - \sqrt{\beta})^2}.$$

The next lemma will play a key role in the proof of Proposition 2, it is similar to Lemma 8 in [1] but it deals with almost balanced communities, this is crucial to our analysis.

Lemma 1 Suppose $\alpha > \beta > 0$. Let X and Y be two independent random variables with $X \sim \text{Binom}(K_1, \alpha \frac{\log n}{n})$ and $Y \sim \text{Binom}(K_2, \beta \frac{\log n}{n})$, where $K_1 = \frac{n\gamma}{2} + o(n)$ and $K_2 = \frac{n\gamma}{2} + o(n)$ as $n \rightarrow \infty$. Then,

$$\mathbb{P}(X - Y \leq 0) \leq n^{-((\alpha+\beta)^{\frac{\gamma}{2}} - \gamma\sqrt{\alpha\beta}) + o(1)}.$$

We present a simple and direct proof of this lemma.

Proof Let $\varepsilon > 0$. We proceed with the Laplace transform method, for all $t \geq 0$ we write

$$\mathbb{P}(X - Y \leq 0) \leq \mathbb{P}(X - Y \leq \varepsilon) \leq e^{t\varepsilon} \mathbb{E}e^{-t(X-Y)} := e^{-\psi(t)}, \tag{3}$$

where $\psi(t) := -t\varepsilon - \log \mathbb{E}e^{-t(X-Y)}$. Now we use the fact that the function $\psi(t)$ is additive for sums of independent random variables together with the formula for the moment generating function of a binomial distribution (Example 3.32 in [12])

$$\log \mathbb{E}e^{-t(X-Y)} = K_1 \log(1 - p(1 - e^{-t})) + K_2 \log(1 - q(1 - e^t)),$$

where $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$. Using the elementary inequality, $\log(1 - x) \leq -x$, valid for all $0 \leq x \leq 1$, we get

$$\psi(t) \geq -\varepsilon t + K_1 p(1 - e^{-t}) + K_2 q(1 - e^t).$$

We pick $t^* = \log((2K_2q)^{-1}(-\varepsilon + \sqrt{\varepsilon^2 + 4K_1K_2pq}))$ in order to optimize the right hand side. The second term in the right hand side becomes

$$K_1 p(1 - e^{-t^*}) = K_1 p \left(1 - \frac{2K_2q}{-\varepsilon + \sqrt{\varepsilon^2 + 4K_1K_2pq}} \right).$$

We are interested in the behaviour of $\psi(t^*)$ when $\varepsilon \rightarrow 0^+$, so we take the limit both sides in the equality above

$$\lim_{\varepsilon \rightarrow 0^+} K_1 p(1 - e^{-t^*}) = K_1 p - \sqrt{K_1 K_2 p q}.$$

Similarly, we get

$$\lim_{\varepsilon \rightarrow 0^+} K_2 q(1 - e^{t^*}) = K_2 q - \sqrt{K_1 K_2 p q}.$$

Now we can take the limit as $\varepsilon \rightarrow 0^+$ in inequality 3 to obtain

$$\mathbb{P}(X - Y \leq 0) \leq e^{\lim_{\varepsilon \rightarrow 0^+} -\psi(t^*)} \leq e^{-(K_1 p + K_2 q - 2\sqrt{K_1 K_2 p q})}.$$

Recall that $K_1 = \frac{n\gamma}{2} + o(n)$, $K_2 = \frac{n\gamma}{2} + o(n)$, $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$. Then,

$$\mathbb{P}(X - Y \leq 0) \leq e^{-\log(n)(\gamma \frac{\alpha+\beta}{2} - \gamma \sqrt{\alpha\beta} + o(1))}.$$

□

We end this section with the proof of Proposition 2.

Proof We denote the success event by \mathcal{E} , i.e, the event that the communities are recovered and we condition on the event that V^\natural has been drawn. By union bound we can write,

$$\mathbb{P}(\mathcal{E}^c \mid V^\natural) \leq P_1 + P_2.$$

Here $P_1 := \sum_{v \in S_1} \mathbb{1}_{\{v \in V(G) \setminus V^\natural\}} \mathbb{P}(e(v, R_1) - e(v, R_2) \leq 0)$ and P_2 is defined analogously.

Observe that now the probability in the right hand side of P_1 is equal to

$$\mathbb{P}\left(\sum_{j=1}^{K_1} B_j^{(p)} - \sum_{j=1}^{K_2} B_j^{(q)} \leq 0\right),$$

where $K_i = |R_i|$ and for all j , the random variables $B_j^p \sim \text{Ber}(p)$ and $B_j^q \sim \text{Ber}(q)$ are all independent. We set $X := \sum_{j=1}^{K_1} B_j^{(p)} \sim \text{Binom}(K_1, \alpha \frac{\log n}{n})$ and $Y := \sum_{j=1}^{K_2} B_j^{(p)} \sim \text{Binom}(K_2, \beta \frac{\log n}{n})$. In order to apply Lemma 1, we denote the event in which both K_1 and K_2 lie in the interval $\frac{n\gamma}{2}(1 \pm \frac{1}{\sqrt{\log n}})$ by \mathcal{A} . So we can bound P_1 by

$$P_1 \leq \sum_{v \in S_1} \mathbb{1}_{\{v \in V(G) \setminus V^\natural\}} (\mathbb{1}_{\{\mathcal{A}^c\}} + \mathbb{1}_{\{\mathcal{A}\}} \mathbb{P}(X - Y \leq 0 \mid \mathcal{A})).$$

We use the crude bound $\mathbb{1}_{\{v \in V(G) \setminus V^\natural\}} \leq 1$ and write

$$P_1 \leq \frac{n}{2} (\mathbb{1}_{\{\mathcal{A}^c\}} + \mathbb{1}_{\{\mathcal{A}\}} \mathbb{P}(X - Y \leq 0 \mid \mathcal{A})).$$

It is easy to see that the same bound holds for P_2 , so

$$\mathbb{P}(\mathcal{E}^c \mid V^\natural) \leq P_1 + P_2 \leq n(\mathbb{1}_{\{\mathcal{A}^c\}} + \mathbb{1}_{\{\mathcal{A}\}} \mathbb{P}(X - Y \leq 0 \mid \mathcal{A})).$$

We take the expectation with respect to V^\natural both sides to obtain,

$$\mathbb{P}(\mathcal{E}^c) \leq n(\mathbb{P}(\mathcal{A}^c) + \mathbb{E}_{V^\natural} \mathbb{P}(X - Y \leq 0 \mid \mathcal{A})). \tag{4}$$

By Chernoff’s small deviation inequality (Exercise 2.3.5 [11]), there is an absolute constant $c > 0$ such that

$$\mathbb{P}(\mathcal{A}^c) \leq 2\mathbb{P}\left(\left|K_1 - \frac{n\gamma}{2}\right| > \frac{n\gamma}{2\sqrt{\log n}}\right) \leq 2e^{-c\frac{\gamma n}{\log n}} = o\left(\frac{1}{n}\right). \tag{5}$$

By Lemma 1,

$$\mathbb{E}_{V^\natural} \mathbb{P}(X - Y \leq 0 \mid \mathcal{A}) \leq n^{-((\alpha+\beta)\frac{\gamma}{2} - \gamma\sqrt{\alpha\beta}) + o(1)}. \tag{6}$$

By the assumption on γ , $(\alpha + \beta - 2\sqrt{\alpha\beta})\frac{\gamma}{2} > 1$. Therefore, there exists an $\varepsilon > 0$ such that

$$\mathbb{P}(X - Y \leq 0 \mid \mathcal{A}) \leq n^{-1-\varepsilon+o(1)} = o\left(\frac{1}{n}\right).$$

Then we combine inequalities 5 and 6 with inequality 4 to complete the proof. \square

3 Exact recovery in the subsampled nodes

In the sampling procedure in Step 1, the unknown communities $S_1 \cap V^\natural$ and $S_2 \cap V^\natural$ are no longer guaranteed to be balanced, therefore we cannot directly use the optimization program (2) because the maximum likelihood estimator is no longer (1). However, thanks to the authors in [8], similar semidefinite programs can be used to handle this case. We follow the approach in [8].

To begin with, it is straightforward to see that if the communities have sizes K and $n - K$, the maximum likelihood estimator becomes

$$\begin{aligned} & \max_x x^T Ax \\ & \text{s.t. } \mathbf{1}^T x = (2K - n) \\ & \quad x \in \{\pm 1\}^n \end{aligned} \tag{7}$$

Therefore we can relax the problem in the same as before, we set $X := xx^T$ and write

$$\begin{aligned} \max_{X \in \mathbb{R}^{n \times n}} \quad & \text{Tr}(AX) \\ \text{s.t.} \quad & X_{ii} = 1 \\ & X \succeq 0 \\ & \text{Tr}(X\mathbf{J}) = (2K - n)^2 \end{aligned} \quad (8)$$

We should remark that the formulation (8) requires the knowledge of the sizes of the communities. To overcome this problem, we consider a Lagrangian formulation

$$\begin{aligned} \max_{X \in \mathbb{R}^{n \times n}} \quad & \text{Tr}(AX) - \lambda^* \text{Tr}(X\mathbf{J}) \\ \text{s.t.} \quad & X_{ii} = 1 \\ & X \succeq 0 \end{aligned} \quad (9)$$

The intuition is that the Lagrange multiplier λ^* adjusts the sizes of the communities. An important insight from [8] is the following: There exists a value of λ^* that works for all values K , so the optimization program (9) can be used to recover unbalanced communities with unknown sizes. Indeed, the following proposition reflects it. We use the notation $G \sim \mathcal{G}(n_1, n_2, p, q)$ to denote a random graph drawn exactly in the same way as before with the exception that now the planted communities have sizes n_1 and n_2 satisfying $n_1 + n_2 = n$ but n_1 is not necessarily equal to n_2 .

Proposition 3 [8] *Let $G \sim \mathcal{G}(K, n - K, p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$ satisfying $p > q$. Then, for $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, the semidefinite program (9) with $\lambda^* = \left(\frac{\alpha - \beta}{\log \alpha - \log \beta} \right) \frac{\log n}{n}$ exactly recovers the communities with probability $1 - Cn^{-c(\frac{1}{2}(\sqrt{\alpha} - \sqrt{\beta})^2 - 1)}$, where $C, c > 0$ are absolute constants.*

4 Main theorem

We shall proceed to the main result of this paper. We combine the ideas in Sects. 2 and 3 to establish a complete analysis of the sketching procedure.

Theorem 4 (Main result) *Let $G \sim \mathcal{G}(n; p, q)$ with planted communities $\{S_1, S_2\}$ and with $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$ satisfying $p > q$. Draw a vertex set V^\square at random by sampling each node of the graph G independently at random with probability γ . Denote, for $i \in \{1, 2\}$, \hat{R}_i to be the maximum likelihood estimators of $R_i = S_i \cap V^\square$ obtained by running the semidefinite program 9 with the input matrix A being the adjacency matrix of the graph $H \subset G$ induced by V^\square and the parameter λ^* chosen as follows: In the event that $|V^\square| \geq 2$, set $\lambda^* = \frac{\alpha_H - \beta_H}{\log \alpha_H - \log \beta_H} \frac{\log |V^\square|}{|V^\square|}$, where $\alpha_H := \frac{p|V^\square|}{\log |V^\square|}$*

and $\beta_H := \frac{q|V^\natural|}{\log|V^\natural|}$, otherwise set $\lambda^* = 0$. Now take

$$\begin{aligned} \hat{S}_1 &= \hat{R}_1 \cup \{v \in V(G) \setminus V^\natural : e(v, \hat{R}_1) > e(v, \hat{R}_2)\}. \\ \hat{S}_2 &= \hat{R}_2 \cup \{v \in V(G) \setminus V^\natural : e(v, \hat{R}_2) > e(v, \hat{R}_1)\}. \end{aligned}$$

Then there exists absolute constants $C, c > 0$ such that, with probability $1 - Cn^{-c((\alpha+\beta)^{\frac{\gamma}{2}} - \gamma\sqrt{\alpha\beta} - 1)}$, $(\hat{S}_1, \hat{S}_2) = (S_1, S_2)$. In particular, with probability $1 - o(1)$, $(\hat{S}_1, \hat{S}_2) = (S_1, S_2)$ as long as

$$\gamma > \frac{2}{(\sqrt{\alpha} - \sqrt{\beta})^2}.$$

Proof Observe that after sampling the vertex set $V(G)$ of the graph, the induced subgraph $H \subset G$ is a random graph with law $H \sim \mathcal{G}(S_1 \cap V^\natural, S_2 \cap V^\natural, p, q)$. We claim that there exists a λ^* such that the optimization program 9 recovers both communities $S_1 \cap V^\natural$ and $S_2 \cap V^\natural$ with the desired probability. The proof of the theorem easily follows from the claim by applying Proposition 2 and union bound.

Now, we proceed to prove the claim. In order to apply Proposition 3 we need to check that, with sufficiently large probability,

$$\sqrt{\alpha_H} - \sqrt{\beta_H} > \sqrt{2}, \tag{10}$$

where $\alpha_H := \frac{p|V^\natural|}{\log|V^\natural|}$ and $\beta_H := \frac{q|V^\natural|}{\log|V^\natural|}$ if $|V^\natural| \geq 2$ and zero otherwise. Recall, by definition, $p = \alpha \frac{\log n}{n}$ and $q = \beta \frac{\log n}{n}$. The degenerate event $|V^\natural| \leq 1$ (empty set or single vertex) occurs with exponentially small probability. Indeed, observe $|V^\natural|$ is a sum of n i.i.d random variables with Bernoulli distribution with mean γ , so

$$\mathbb{P}(|V^\natural| \leq 1) = (1 - \gamma)^n + n(1 - \gamma)^{n-1}\gamma \leq 2e^{-\gamma(n-1) + \log n},$$

and

$$\mathbb{P}\left(\alpha_H = \alpha \frac{|V^\natural| \log n}{n \log |V^\natural|} \cap |V^\natural| \geq 2\right) = 1 - 2e^{-\gamma(n-1) + \log n}.$$

An analogous fact holds for β_H , so the event

$$\left\{ \sqrt{\alpha_H} - \sqrt{\beta_H} = \sqrt{\frac{|V^\natural| \log n}{n \log |V^\natural|}} (\sqrt{\alpha} - \sqrt{\beta}) \right\} \cap \{|V^\natural| \geq 2\},$$

occurs with exponentially large probability. Since $\frac{\log n}{\log |V^\natural|} \geq 1$ (when the quotient makes sense), it is enough to prove that, with the desired probability,

$$\left(\sqrt{\frac{|V^\natural|}{n}} \right) (\sqrt{\alpha} - \sqrt{\beta}) > \sqrt{2}.$$

By assumption, there exists a $\delta > 0$ such that $\sqrt{\alpha} - \sqrt{\beta} \geq \sqrt{\frac{2}{\gamma}}(1 + \delta)$ and by the small Chernoff deviation inequality, for every $\varepsilon > 0$, $\mathbb{P}\left(\frac{|V^\natural|}{n} \geq \gamma - \varepsilon\right) \geq 1 - 2e^{-c\varepsilon^2 n \gamma^3}$. Putting these three facts together, we obtain, for every $\varepsilon > 0$,

$$\left(\sqrt{\frac{|V^\natural| \log n}{n \log |V^\natural|}} \right) (\sqrt{\alpha} - \sqrt{\beta}) \geq \sqrt{2}(1 + \delta) \sqrt{1 - \frac{\varepsilon}{\gamma}},$$

with exponentially large probability. We choose $\varepsilon > 0$ small enough to guarantee that $(1 + \delta) \sqrt{1 - \frac{\varepsilon}{\gamma}} > \sqrt{1 + \delta}$ and then inequality (10) is satisfied with the desired probability. The claim now follows from Proposition 3. \square

Acknowledgements The authors would like to thank Dustin Mixon, Kaiying Xie and Nikita Zhivotovsky for helpful discussions. The authors would also like to thank anonymous referees for valuable comments that improved the manuscript.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abbe, E., Bandeira, A.S., Hall, G.: Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* **62**(1), 471–487 (2016)
2. Abbe, E.: Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* **18**(1), 6446–6531 (2017)
3. Alizadeh, F.: Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Optim.* **5**(1), 13–51 (1995)
4. Bandeira, A.S.: Random Laplacian matrices and convex relaxations. *Found. Comput. Math.* **18**(2), 345–379 (2018)
5. Bluhm, A., França, D.S.: Dimensionality reduction of SDPs through sketching. *Linear Algebra Appl.* **563**, 461–475 (2019)
6. Garey, M.R., Johnson, D.S.: *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, San Francisco (1979)

7. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM (JACM)* **42**(6), 1115–1145 (1995)
8. Hajek, B., Wu, Y., Xu, J.: Achieving exact cluster recovery threshold via semidefinite programming: extensions. *IEEE Trans. Inform. Theory* **62**(10), 5918–5937 (2016)
9. Mixon, D.G., Xie, K.: Sketching semidefinite programs for faster clustering. *IEEE Trans. Inform. Theory* **67**(10), 6832–6840 (2021)
10. Mossel, E., Neeman, J., Sly, A.: Consistency thresholds for the planted bisection model. In: *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, pp. 69–75 (2015)
11. Vershynin, R.: *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge (2018)
12. Wasserman, L.: *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York (2013)
13. Woodruff, D.P.: Sketching as a tool for numerical linear algebra. In: *Foundations and Trends® in Theoretical Computer Science* 10(1–2), 1–157 (2014)
14. Yurtsever, A., Udell, M., Tropp, J., Cevher, V.: Sketchy decisions: convex low-rank matrix optimization with optimal storage. In: *Artificial Intelligence and Statistics*, pp. 1188–1196 (2017)