Lecture 14: The Stochastic Block Model spectral & semidefinite programming approaches

Foundations of Data Science: Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

27 September, 2023

Problem setup

Weak recovery & spectral relaxation

Exact recovery & SDP relaxation

Stochastic Block Model

- consider a random graph model which induces a clustering structure on the generated graph instance
- fix $n \in \mathbb{N}_+$, and consider two sets $\mathcal{C}_1$ and $\mathcal{C}_2$, each of size $m = \frac{n}{2}$

$$|\mathcal{C}_1| = |\mathcal{C}_2| = m$$

- for each pair of nodes (i, j)
  - $(i, j) \in E(G)$ with prob p, if i and j are in the **same** cluster
  - $(i, j) \notin E(G)$ with prob q, if i and j are in **different** clusters

$$A_{ij} = \begin{cases} 1 & \text{if i and j are in the same cluster} \\ 0 & \text{if i and j are in different clusters} \end{cases} \tag{1}$$

- each edge is drawn independently
- typically, it is the case that $p > q$. Note the special cases
  - $p = 1$ and $q = 0$ renders the problem trivial
  - $p = q = \frac{1}{2}$ make it impossible to recover the two clusters
  - the case $p < q$ can be treated symmetrically (one can also consider the complement graph)
- research question: for which values of p and q is it possible to recover the underlying partition? (i.e, the two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$) - or at least do so with high probability.

Stochastic Block Model

The case of $k = 2$ equally-sized communities

$$\begin{pmatrix} p & q \\ q & p \end{pmatrix}$$

(a) Scrambled graph

(b) Clustered graph and color-coded

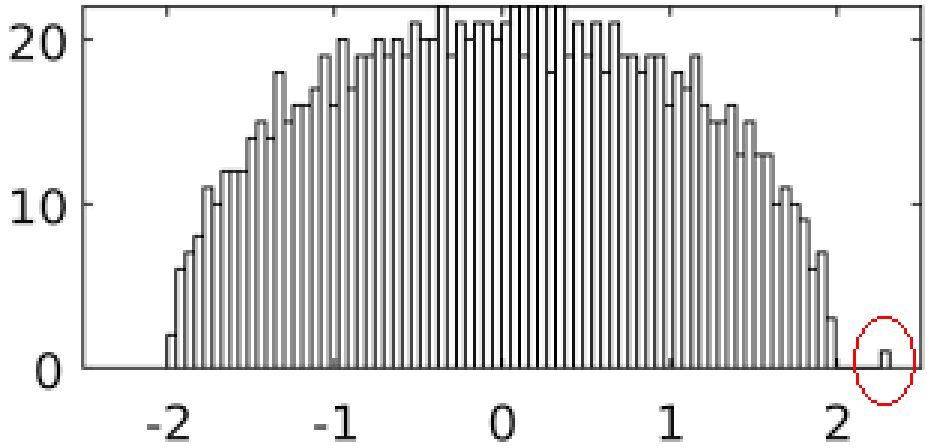
Figure: A graph instance generated form the SBM with $n = 600$ nodes and $k = 2$ communities, with within-cluster probability $p = 6/600$ and across-cluster probability $q = 0.1/600$ (Image source: Emmanuel Abbe).

Spiked Wigner Model

- we observe

$$Y = \lambda x x^T + \frac{1}{\sqrt{n}} W \tag{2}$$

- where W is a $n \times n$ random symmetric matrix with entries drawn i.i.d. (up to symmetry) from a fixed distribution of mean 0 and variance 1
- rank-1 perturbation of a Wigner matrix
- the top eigenvalue of Y separates from the semicircular bulk when $\lambda > 1$ (Péché, 2006; Féral and Péché 2007)



Spiked (Gaussian) Wishart Model

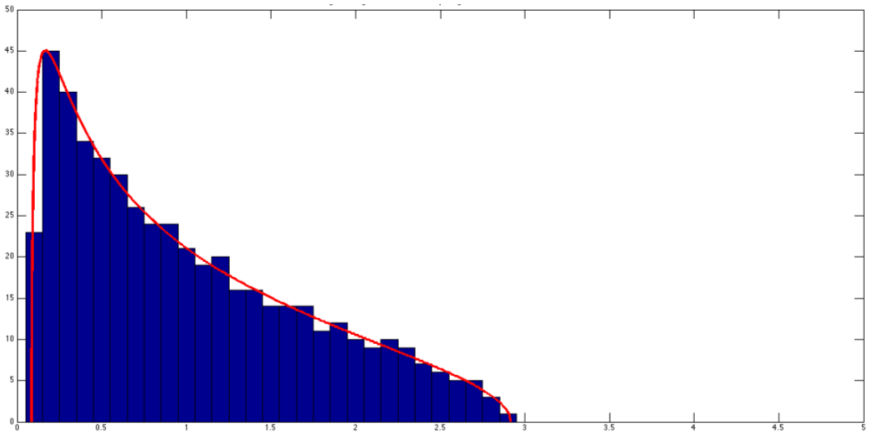
- we observe the sample covariance (across T observations)

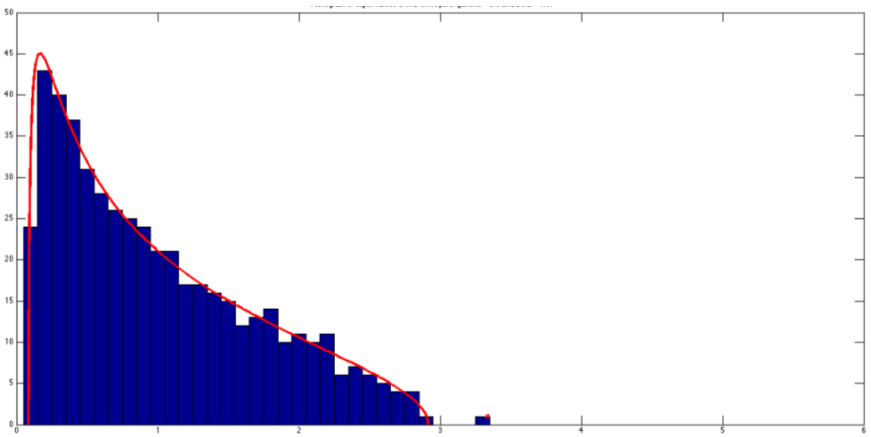
$$Y = \frac{1}{T} X X^\top \tag{3}$$

- with X an $n \times T$ matrix with columns drawn iid from $\mathcal{N}(0, I_n + \beta x x^\top)$ in the high-dim. setting where the sample count T and dimension n scale proportionally as $\frac{n}{T} \to \gamma$, & $\beta \in [-1, \infty]$
- Baik, Ben Arous and Péché (2005) showed that, when

$$\beta > \sqrt{\gamma},$$

an isolated eigenvalue emerges ("pops out") from the bulk of the Marchenko-Pastur distribution.





Intuition from the spike models

- work towards a spectral relaxation for the clustering task
- consider the adjacency matrix of the graph G

$$A_{ij} = \begin{cases} 1 & \text{if}(i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

- note that A is a random matrix
- keeping the clustering task in mind, a natural objective to consider is the quadratic form

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t.} \quad & x_i = \pm 1, \forall i = 1, \ldots, n, \\ & \sum_j x_j = 0 \end{aligned}$$

- ideally, the target solution is such that x takes value $+1$ in one cluster, and -1 in the other clusters (eg., $x_i = +1$ if $i \in \mathcal{C}_1$, and $x_i = -1$ if $i \in \mathcal{C}_2$).

Intuition from the spike model analysis (cont)

- relaxing the condition

$$x_i = \pm 1, \forall i = 1, \ldots, n$$

to

$$||x||_2^2 = n$$

leads to the spectral relaxation method

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t.} \quad & ||x||_2 = \sqrt{n}, \\ & \mathbf{1}^T x = 0 \end{aligned}$$

- the solution that maximizes the quadratic form $x^T Ax$ is given by the top eigenvector of the projection of A onto the orthogonal of the all-ones vector $\mathbf{1}$

Expected adjacency matrix

- the adjacency matrix A is a random matrix with expectation

$$\mathbb{E}[A] = \begin{cases} p & \text{if}(i,j) \in E(G) \\ q & \text{otherwise} \end{cases} \tag{5}$$

- let w denote the vector taking values
  - $+1$ on nodes in cluster $\mathcal{C}_1$,
  - -1 on nodes in cluster $\mathcal{C}_2$.

  WLOG we can assume that

$$w = (\underbrace{1, \ldots, 1}_{n/2}, \underbrace{-1, \ldots, -1}_{n/2})^T \in \mathbb{R}^n \tag{6}$$

  corresponding to the “ground truth” or the ”planted clusters” we seek to recover
- via simple algebraic manipulations we can write

$$\mathbb{E}[A] = \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2}ww^T \tag{7}$$

- which can be further written as

$$A = (A - \mathbb{E}[A]) + \mathbb{E}[A]$$

$$A = (A - \mathbb{E}[A]) \quad + \quad \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2}ww^T \tag{8}$$

Rank-1 perturbation

- to remove the term $\frac{p+q}{2}\mathbf{1}\mathbf{1}^T$, we consider the following rank-1 update to A, and define the random matrix $\mathcal{A}$ given by

$$\mathcal{A} = A - \frac{p+q}{2}\mathbf{1}\mathbf{1}^T \tag{9}$$

- by considering the expectation of $\mathcal{A}$, from (8), one arrives at

$$\mathcal{A} = (\mathcal{A} - \mathbb{E}\mathcal{A}) - \frac{p-q}{2}ww^T \tag{10}$$

- the decomposition renders $\mathcal{A}$ as a **superposition** of a random matrix whose expected value is 0, and a rank-1 matrix

$$\mathcal{A} = W - \frac{p-q}{2}ww^T \tag{11}$$

- i.e, $\mathcal{A}$ is a rank-1 perturbation of a random matrix

$$\mathcal{A} = W + \lambda vv^T \tag{12}$$

where

- $W = (\mathcal{A} - \mathbb{E}\mathcal{A})$
- $\lambda vv^T = \frac{p-q}{2}n\left(\frac{w}{\sqrt{n}}\right)\left(\frac{w}{\sqrt{n}}\right)^T$

Rank-1 perturbation (cont)

- random matrix theory tells us that for large enough λ
  - the top eigenvalue associated to λ will "pop" outside the distribution of the eigenvalues of W
  - its corresponding eigenvector will have a non-trivial correlation with the true signal w
- one can further rewrite the previous optimization as

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i,j} \mathcal{A}_{ij} x_i x_j \\ \text{s.t.} \quad & ||x||_2 = \sqrt{n}, \\ & \mathbf{1}^T x = 0 \end{aligned}$$

- since we have subtracted from A a scalar multiple of $\mathbf{1}\mathbf{1}^T$, this allows for dropping the constraint $\mathbf{1}^T x = 0$, leading to

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i,j} \mathcal{A}_{ij} x_i x_j \\ \text{s.t.} \quad & ||x||_2 = \sqrt{n} \end{aligned}$$

whose solution is simply given by the top eigenvector of $\mathcal{A}$.

Informal argument from RMTX

- if $W \stackrel{\text{def}}{=} (\mathcal{A} - \mathbb{E}\mathcal{A})$ was a Wigner matrix with i.i.d entries, zero mean and variance σ^2, then
  - its empirical spectral density would follow the semicircle law
  - the bulk of the distribution supported in $[-2\sigma\sqrt{n}, 2\sigma\sqrt{n}]$
  - RMTX tells us that we would expect the top eigenvector of $\mathcal{A}$ to correlate with the ground truth w as soon as

$$\frac{p-q}{2}n > \frac{2\sigma\sqrt{n}}{2} \tag{13}$$

- however, W is not a Wigner matrix in general; half of its entries have variance $p(1-p)$, and half have $q(1-q)$
- so if we were to plugin $\sigma^2 = \frac{p(1-p)+q(1-q)}{2}$ in (13), it would hint that the top eigenvector of $\mathcal{A}$ correlates with w as soon as

$$\frac{p-q}{2} > \frac{1}{\sqrt{n}}\sqrt{\frac{p(1-p)+q(1-q)}{2}} \tag{14}$$

- for the special case $q = 1-p$ (thus $p = 1-q$)
  - the entries of W have the same variance
  - would still imply the non-trivial result that $p-q$ only needs to be around $\frac{1}{\sqrt{n}}$ in order for the top eigenvector to correlate with the ground truth w (impressive in itself!)

The case of sparse graphs

- in many real-world applications, such as social networks, the average degree of each node is constant
- consider for example the case when $p = \frac{a}{n}$ and $q = \frac{b}{n}$, for some fixed constants a and b
- following the previous (non-rigorous) line of thought, the following was proposed as a conjecture by Decelle et al. (2011); partial/weak recovery is feasible w.p. $1 - o(1)$ whenever

$$(a-b)^2 > 2(a+b) \tag{15}$$

- it attracted a lot of attention, and was ultimately proved in a series of works by Mossel et al. (2014) and independently by Massoulie (2014), by studying variants of belief propagation using techniques from statistical physics
- below the connectivity threshold $(a-b)^2 < 2(a+b)$, any estimator fails with probability $1 - o(1)$
- comprehensive survey
  - Emmanuel Abbe, *Community detection and stochastic block models*. Foundations and Trends in Communications and Information Theory, 14(1-2):1– 162, 2018

Dense versus sparse graphs

- given graph G concentrates about its expectation if A is close to its expectation $\mathbb{E}[A]$ in some natural matrix norm
- interpret the expectation of G as the weighted graph with adjacency matrix $\mathbb{E}[A]$
- various matrix norms could be of interest; in previous slides we have looked at the spectral norm $\max|\lambda_i|$
- often, the question of interest is estimating some features of the probability matrix Π_{ij} from random graphs drawn from $G(n, \Pi_{ij})$
- concentration of the adjacency and Laplacian matrix around their expectations, when it holds, guarantees recovery of such features
- denote the expected degree by $d = pn$
- dense graphs: with high probability

$$||A - \mathbb{E}[A]|| = 2\sqrt{d}(1 + o(1)) \quad \text{if } d \gg log^4 n \tag{16}$$

- as $||\mathbb{E}[A]|| = d$ the typical deviation behaves like the square root of the magnitude of expectation (like in other classical results of probability theory) $\Rightarrow$ *dense random graphs concentrate well*
- lower bound can be relaxed all the way down to $d = \Omega(\log n)$

Combinatorics, Probability and Computing (2003) 12, 61–72.
DOI: 10.1017/S0963548302005424 Printed in the United Kingdom

The Largest Eigenvalue of Sparse Random Graphs

MICHAEL KRIVELEVICH[1]† and BENNY SUDAKOV[2]‡

[1] Department of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
(e-mail: krivelev@math.tau.ac.il)

[2] Department of Mathematics, Princeton University, Princeton, NJ 08544, USA
and
Institute for Advanced Study, Princeton, NJ 08540, USA
(e-mail: bsudakov@math.princeton.edu)

Received 7 June 2001; revised 15 July 2002

We prove that, for all values of the edge probability $p(n)$, the largest eigenvalue of the random graph $G(n,p)$ satisfies almost surely $\lambda_1(G) = (1+o(1))\max\{\sqrt{\Delta}, np\}$, where Δ is the maximum degree of G, and the $o(1)$ term tends to zero as $\max\{\sqrt{\Delta}, np\}$ tends to infinity.

Dense versus sparse graphs

- problem: sparse graphs do not concentrate; in the sparse regime, for bounded expected degree d, concentration breaks down
- one can show that a random graph from $G(n,p)$ satisfies w.h.p

$$||A|| = (1+o(1))\sqrt{d(A)} = (1+o(1))\sqrt{\frac{\log n}{\log\log n}}, \quad \text{if } d = O(1) \tag{17}$$

- $d(A) :=$ maximal degree of the graph (a random quantity)
- $||A|| \gg ||\mathbb{E}[A]|| = d \Rightarrow$ *sparse random graphs do not concentrate*
- what exactly makes $||A||$ abnormally large in the sparse regime?
- vertices with too high degrees!
  - in the dense case $d \gg \log n$, all vertices typically have approximately the same degrees $(1+o(1))d$
  - no longer the case in the sparser regime $d \ll \log n$; the degrees do not cluster tightly about the same value anymore.
  - there are vertices with too high degrees; even a single high-degree vertex can blow up the norm of the adjacency matrix
  - since the norm of A is bounded below by the Euclidean norm of each of its rows, we have $||A|| \geq \sqrt{d(A)}$
- calls for regularization techniques; for eg, $A := A + \tau \mathbf{1}\mathbf{1}^T$, $\tau \in \mathbb{R}$

The case of $k \geq 3$ communities

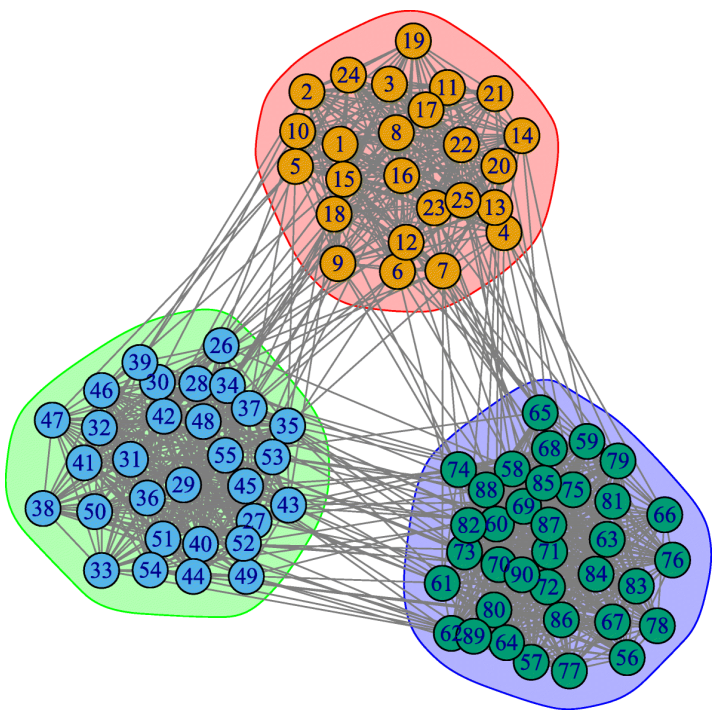
- balanced symmetric stochastic block model: k communities of equal size
- in the sparse regime of
  - within-cluster probability $p = \frac{a}{n}$
  - across-cluster probability $q = \frac{b}{n}$

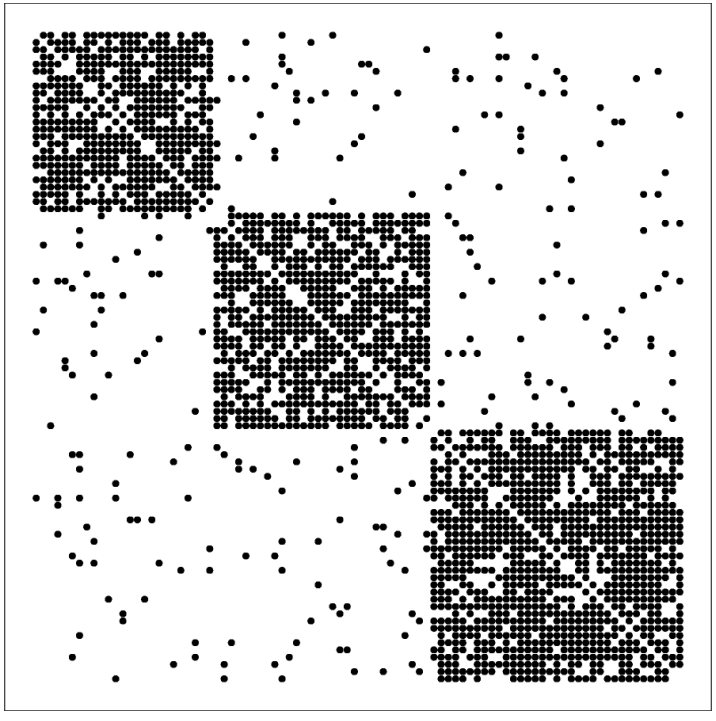
$$\begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}$$

- conjectured to have a statistical-to-computational gap, meaning:
  - there is a range of parameters a and b such that the problem of partial recovery is statistically or information-theoretically possible
  - but there does not exist a polynomial-time algorithm for this
- insights driven by tools from the statistical physics literature.

Stochastic block modeling ($k = 3$) - simulation

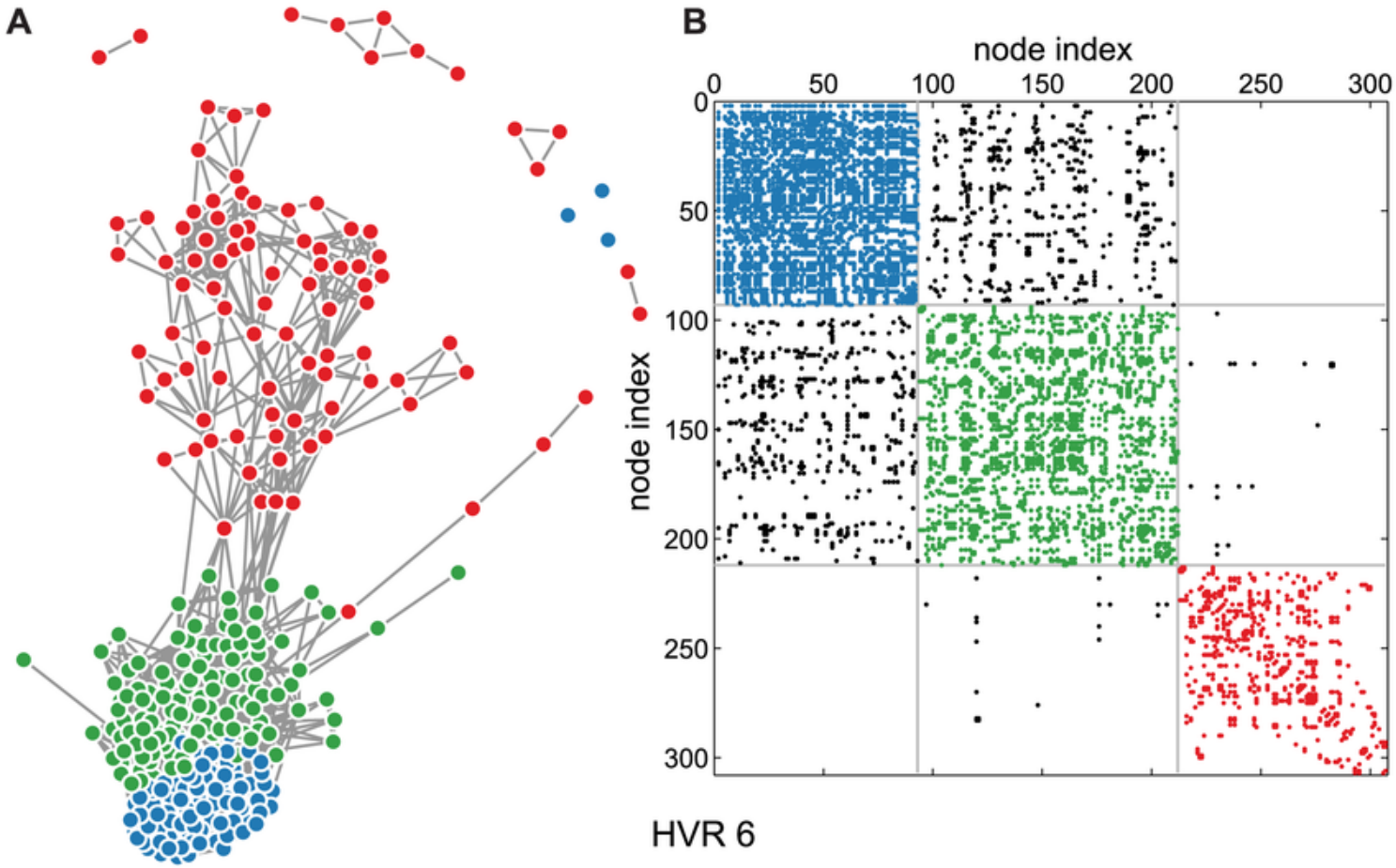
$$\begin{pmatrix} 0.8 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.8 \end{pmatrix}$$





Lee, C., Wilkinson, D.J. A review of stochastic block models and extensions for graph clustering. Appl Netw Sci 4, 122 (2019).
https://doi.org/10.1007/s41109-019-0232-2

Stochastic block modeling in real data ($k = 3$)



Larremore, Daniel B., Aaron Clauset, and Caroline O. Buckee. ”A network approach to analyzing highly recombinant malaria parasite genes.” PLoS Comput Biol 9.10 (2013): e1003268.

What about exact recovery? ($k = 2$)

- moving beyond partial/weak recovery (of simply having an estimate that correlates with the true labels)
- exact recovery entails **recovery of cluster membership of each and every single node correctly**
- if the inner cluster probability is of order $p = \frac{a}{n}$, then
  - the graph would have isolated nodes (of degree 0)
  - impossible to recover the cluster membership of each node
  - same holds true for $p << \frac{2 \log n}{n}$
- focus on following regime (for some constants , $\beta > 0$)

$$p = \frac{\alpha \log(n)}{n} \quad \text{and} \quad q = \frac{\beta \log(n)}{n} \tag{18}$$

- recall the minimum bisection objective

$$\max_{x \in \mathbb{R}^n} \sum_{i,j} A_{ij} x_i x_j$$

$$\text{s.t.} \quad x_i = \pm 1, \forall i \quad \text{and} \quad x^T \mathbf{1} = 0$$

- if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, then the above recovers the true partition w.h.p (information-theoretic impossible w.h.p if $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$).

Semidefinite programming (SDP)

- SDP is a branch of convex programming, where the optimization (of a linear objective function) takes place over the cone of positive semidefinite matrices
- eg, consider a linear function of X

$$C \cdot X := \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij} \tag{19}$$

Recall $\mathrm{Tr}(CX) = \sum_{i,j=1}^{n} C_{ij} X_{ji} = \mathrm{Tr}(XC)$

- in an SDP, the variable is the matrix X, but it might be helpful to think of X as an array of n^2 numbers

$$\begin{aligned} \min \quad & C \cdot X \\ \text{s.t.} \quad & A_i \cdot X = b_i, \forall i = 1, \ldots, n, \\ & X \succeq 0 \end{aligned} \tag{20}$$

Relaxation via semidefinite programming (SDP)

$$\max_{x \in \mathbb{R}^n} \quad \sum_{i,j} A_{ij} x_i x_j$$

$$\text{s.t.} \qquad x_i = \pm 1, \forall i \quad \text{and} \quad x^T \mathbf{1} = 0$$

- if we remove the constraint $x^T \mathbf{1} = 0$, then the optimal solution becomes $x = \mathbf{1}$
- define $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$, leading to

$$B_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } (i,j) \in E(G) \\ -1 & \text{otherwise} \end{cases} \tag{21}$$

- one could verify that the following problem has the same solution as the optimization problem at the top

$$\max_{x \in \mathbb{R}^n} \quad \sum_{i,j} B_{ij} x_i x_j$$

$$\text{s.t.} \qquad x_i = \pm 1, \forall i \quad \text{and} \quad x^T \mathbf{1} = 0$$

Relaxation via semidefinite programming (SDP)

- if we remove the constraint $x^T \mathbf{1} = 0$, then $x = \mathbf{1}$ is no longer the optimal solution
- intuitively, the penalization created by subtracting a multiple of $\mathbf{1}\mathbf{1}^T$ is enough to discourage unbalanced partitions
- next, we aim to solve efficiently

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & \sum_{i,j} B_{ij} x_i x_j \\ \text{s.t.} \quad & x_i = \pm 1, \forall i \end{aligned}$$

- which is in general computational-hard to solve (NP-hard; reduction from *Max-Cut*)
- relax to an easier problem by similar techniques used to approximate the *Max-Cut* problem (**matrix lifting**)
- by writing $X = xx^T$, the above objective becomes

$$\sum_{i,j} B_{ij} x_i x_j = x^T B x = \mathrm{Tr}(x^T B x) = \mathrm{Tr}(B x x^T) = \mathrm{Tr}(BX) \tag{22}$$

- note that the condition $x_i = \pm 1$ implies $X_{ii} = x_i^2 = 1$

Relaxation via semidefinite programming (SDP)

- we can now re-write the previous optimization problem as

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall i, \\ & X = xx^T, \text{ for some } x \in \mathbb{R}^n \end{aligned}$$

- the latter constraint $X = xx^T$, for some $x \in \mathbb{R}^n$ is equivalent to $\mathrm{rank}(X) = 1$ and $X \succeq 0$, leading to the optimization

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall i, \\ & rank(X) = 1, \\ & X \succeq 0 \end{aligned}$$

- since rank constraints are NP-hard, we relax the problem by removing the non-convex rank constraint

Relaxation via semidefinite programming (SDP)

- we arrive at the following SDP which can be solved (up to arbitrary precision) in polynomial time

$$\begin{aligned} \max_{X} \quad & \mathrm{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall i, \\ & X \succeq 0 \end{aligned}$$

- since we had removed the rank-1 constraint, the solution to the above is no longer guaranteed to be rank-1
- recall

$$p = \frac{\alpha \log(n)}{n} \quad \text{and} \quad q = \frac{\beta \log(n)}{n} \tag{23}$$

- one can show that, for some values of α and β, with high prob.
  - the solution to the SDP satisfies the rank-1 constraint
  - and also coincides with $X = ww^T$, where w corresponds to the true partition.
- after X is computed, w is obtained as its top eigenvector.

Recent survey on SBM (2018)

Journal of Machine Learning Research 18 (2018) 1-86 Submitted 9/16; Revised 3/17; Published 4/18

Community Detection and Stochastic Block Models: Recent Developments

Emmanuel Abbe EABBE@PRINCETON.EDU
Program in Applied and Computational Mathematics
and Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: Edoardo M. Airoldi

Recent survey on SBM (2018)

Abstract

The stochastic block model (SBM) is a random graph model with planted clusters. It is widely employed as a canonical model to study clustering and community detection, and provides generally a fertile ground to study the statistical and computational tradeoffs that arise in network and data sciences.

This note surveys the recent developments that establish the fundamental limits for community detection in the SBM, both with respect to information-theoretic and computational thresholds, and for various recovery requirements such as exact, partial and weak recovery (a.k.a., detection). The main results discussed are the phase transitions for exact recovery at the Chernoff-Hellinger threshold, the phase transition for weak recovery at the Kesten-Stigum threshold, the optimal distortion-SNR tradeoff for partial recovery, the learning of the SBM parameters and the gap between information-theoretic and computational thresholds.

The note also covers some of the algorithms developed in the quest of achieving the limits, in particular two-round algorithms via graph-splitting, semi-definite programming, linearized belief propagation, classical and nonbacktracking spectral methods. A few open problems are also discussed.

Keywords: Community detection, clustering, stochastic block models, random graphs, unsupervised learning, spectral algorithms, computational gaps, network data analysis.

Abbe, Emmanuel. ”Community detection and stochastic block models: recent developments.” The Journal of Machine Learning Research 18.1 (2017): 6446-6531.