COMMUNITY DETECTION VIA SEMIDEFINITE RELAXATION

AFONSO S. BANDEIRA

Abstract. Notes for lecture given by the author on November 7, 2014 as part of the special course: "Randomness, Matrices and High Dimensional Problems", at IMPA, Rio de Janeiro, Brazil.

The results presented in these notes are from [1].

1. The problem we will focus on

Let n be an even positive integer. Given two sets of $\frac{n}{2}$ nodes consider the following random graph G: For each pair (i,j) of nodes, (i,j) is an edge of G with probability p if i and j are in the same set and q if they are in different sets. Each edge is drawn independently and $p > q$.

(Think nodes as fans of Fluminense and Flamengo and edges representing friendships, in this model, fans of the same club are more likely to be friends)

For which values of p and q can we recover the partition, with an efficient algorithm, from only looking at the graph G (with high probability)?

2. The interesting regime

If $p \ll \frac{\log n}{n}$ then it is easy to see that each cluster will not be connected (with high probability) and so recovery is not possible. In fact, the interesting regime is when

$$p = \frac{\alpha \log(n)}{n} \text{ and } q = \frac{\beta \log(n)}{n}, \tag{2.1}$$

for constants $\alpha > \beta$.

Let A be the adjacency matrix of G, meaning that

$$A_{ij} = \begin{cases} 1 \text{ if } (i,j) \in E(G) \\ 0 \text{ otherwise.} \end{cases} \tag{2.2}$$

Let $x \in \mathbb{R}^n$ with $x_i = \pm 1$ represent a partition (note there is an ambiguity in the sense that x and $-x$ represent the same partition). Then, if we did not worry about efficiency then our guess (which corresponds to the Maximum Likelihood Estimator) would be the solution of

$$\begin{aligned} \max \ & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t. } & x_i = \pm 1, \forall_i \\ & \sum_j x_j = 0, \end{aligned} \tag{2.3}$$

as this maximizes the number of edges within the clusters minus the number of edges across the clusters (the clusters being each set of the partition given by x).

The author would like to thank IMPA for the hospitality.

In fact, one can show (but will not be the focus of this lecture, see [1] for a proof) that if

$$\frac{\alpha+\beta}{2}-\sqrt{\alpha\beta}>1, \tag{2.4}$$

then, with high probability, (2.3) recovers the true partition. Moreover, if

$$\frac{\alpha+\beta}{2}-\sqrt{\alpha\beta}<1,$$

no algorithm (efficient or not) cannot, with high probability, recover the true partition.

In this lecture we are interested in understanding when the true partition can be recovered, using an efficient algorithm.

3. The algorithm

Note that if we remove the constraint that $\sum_j x_j = 0$ in (2.3) then the optimal solution becomes $x \equiv 1$. Let us define $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$, meaning that

$$B_{ij} = \begin{cases} 0 \text{ if } i = j \\ 1 \text{ if } (i,j) \in E(G) \\ -1 \text{ otherwise.} \end{cases} \tag{3.1}$$

It is clear that the problem

$$\begin{aligned} \max & \sum_{i,j} B_{ij}x_ix_j \\ \text{s.t. } & x_i = \pm 1, \forall_i \\ & \sum_j x_j = 0, \end{aligned} \tag{3.2}$$

has the same solution has (2.3). However, when the constraint is dropped,

$$\begin{aligned} \max & \sum_{i,j} B_{ij}x_ix_j \\ \text{s.t. } & x_i = \pm 1, \forall_i, \end{aligned} \tag{3.3}$$

$x \equiv 1$ is no longer an optimal solution. Intuitively, there is enough "-1" contributions to discourage unbalanced partitions. In fact, (3.3) is the problem we'll set ourselves to solve.

Unfortunately (3.3) is in general NP-hard (one can encode, for example, `Max-Cut` by picking the right B). We will relax it to an easier problem by a technique known as *lifting*. If we write $X = xx^T$ then we can formulate the objective value of (3.3) as

$$\sum_{i,j} B_{ij}x_ix_j = x^TBx = \operatorname{Tr}(x^TBx) = \operatorname{Tr}(Bxx^T) = \operatorname{Tr}(BX),$$

also, the condition $x_i = \pm 1$ can be translated in $X_ii = x_i^2 = 1$. This means that (3.3) is equivalent to

$$\begin{aligned} \max & \operatorname{Tr}(BX) \\ \text{s.t. } & X_{ii} = 1, \forall_i \\ & X = xx^T \text{ for some } x \in \mathbb{R}^n. \end{aligned} \tag{3.4}$$

The fact that $X = xx^T$ for some $x \in \mathbb{R}^n$ is equivalent to $\operatorname{rank}(X) = 1$ and $X \succeq 0$ (X is Positive semidefinite, meaning that it is symmetric and all it's eigenvalues are non-negative). This means that (3.3) is equivalent to

$$\begin{aligned} \max\ & \operatorname{Tr}(BX) \\ \text{s.t. } & X_{ii} = 1, \forall_i \\ & X \succeq 0 \\ & \operatorname{rank}(X) = 1. \end{aligned} \tag{3.5}$$

(*But wait, if it is equivalent then why isn't this problem just as hard?*) — It is just as hard. The idea is that now we relax the problem and remove one constraint, the rank constraint

$$\begin{aligned} \max\ & \operatorname{Tr}(BX) \\ \text{s.t. } & X_{ii} = 1, \forall_i \\ & X \succeq 0. \end{aligned} \tag{3.6}$$

Now, (3.6) is an optimization problem with a linear objective and convex feasibility set, moreover the feasibility set is a nice convex set, this type of problems are known as Semidefinite Programs and can be solved (up to arbitrary precision) in polynomial time [4].

(*But wait, if we removed a constraint then why would the solution of this problem have that particular form?*) — Indeed, it won't in general. What we will show is that, for some values of α and β, with high probability, the solution to (3.6) not only satisfies the rank constraint but it coincides with $X = gg^T$ where g corresponds to the true partition. After computing such X, recovering g from it is trivial (leading eigenvector).

Remark 3.1. There are other types of results that, instead of assuming stochastic input, assume worst-case input and analyze the performance of rounding procedures that graph solutions to the relaxation and "round them" to solution of the original problems. I recommend taking a look at a very nice relaxation for `Max-Cut` in [2]

4. The analysis

WLOG we can assume that $g = (1, ..., 1, -1, ..., -1)^T$, meaning that the true partition corresponds to the first $\frac{n}{2}$ nodes on one side and the other $\frac{n}{2}$ on the other. Nothing changes but makes it easier to think about.

4.1. Some preliminary definitions.

Recall that the degree matrix D of a graph G is a diagonal matrix where each diagonal coefficient D_{ii} corresponds to the number of neighbours of vertex i and that $\lambda_2(M)$ is the second smallest eigenvalue of a symmetric matrix M.

Definition 4.1. Let $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) be the subgraph of G that includes the edges that link two nodes in the same community (resp. in different communities) and A the adjacency matrix of G. We denote by $D_{\mathcal{G}}^+$ (resp. $D_{\mathcal{G}}^-$) the degree matrix of $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) and define the Stochastic Block Model Laplacian to be

$$L_{SBM} = D_{\mathcal{G}}^+ - D_{\mathcal{G}}^- - A$$

4.2. Convex Duality.

A standard technique to show that a candidate solution is the optimal one for a convex problem is to use convex duality.

The dual problem of (3.6) is

$$\begin{aligned} \min\ & \operatorname{Tr}(Z) \\ \text{s.t. } & Z \text{ is diagonal} \\ & Z - B \succeq 0. \end{aligned} \tag{4.1}$$

The objective value of the dual is always larger or equal to the primal. In fact, let X, Z be respectively a feasible point of (3.6) and (4.1), then:

Since $Z - B \succeq 0$ and $X \succeq 0$ then $\operatorname{Tr}[(Z-B)X] \geq 0$ which means that

$$\operatorname{Tr}(Z) - \operatorname{Tr}(BX) = \operatorname{Tr}[(Z-B)X] \geq 0,$$

as we wanted. Recall that we want to show that gg^T the optimal solution of (3.6). Then, if we find Z diagonal, such that $Z - B \succeq 0$ and

$$\operatorname{Tr}[(Z-B)gg^T] = 0, \quad \text{(this condition is known as complementary slackness)}$$

then $X = gg^T$ must be an optimal solution of (3.6). To ensure that gg^T is the unique solution we just have to ensure that the nullspace of $Z - B$ only has dimension 1 (which corresponds to multiples of g). Essentially, if this is the case, then for any other possible solution X one could not satisfy complementary slackness.

This means that if we can find Z such that:

(1) Z is diagonal
(2) $\operatorname{Tr}[(Z-B)gg^T] = 0$
(3) $Z - B \succeq 0$
(4) $\lambda_2(Z-B) > 0$,

then gg^T is the unique optima of (3.6) and so recovery of the true partition is possible (with an efficient algorithm).

Z is known as the dual certificate, or dual witness.

4.3. Building the dual certificate.

The idea to build Z is to construct it to satisfy properties (1) and (2) and try to show that it satisfies (3) and (4) using concentration.

If indeed $Z - B \succeq 0$ then (2) becomes equivalent to $(Z-B)g = 0$. This means that we need to construct Z such that $Z_{ii} = \frac{1}{g_i} B[i,:]g$. Since $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$ we have

$$Z_{ii} = \frac{1}{g_i}(2A - (\mathbf{1}\mathbf{1}^T - I))[i,:]g = 2\frac{1}{g_i}Ag + 1,$$

meaning that

$$Z = 2(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-) + I.$$

This means that

$$Z - B = 2(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-) - I - \left[2A - (\mathbf{1}\mathbf{1}^T - I)\right] = 2L_{SBM} + 11^T$$

It trivially follows (by construction) that

$$(Z-B)g = 0.$$

This means that

Lemma 4.2. *If*

$$\lambda_2(2L_{SBM} + 11^T) > 0, \tag{4.2}$$

then the relaxation recovers the true partition.

Note that $2L_{SBM} + 11^T$ is a random matrix and so, now, this is “only” an exercise in random matrix theory.

4.4. **Matrix Concentration.** And easy calculation gives

$$\mathbf{E}\left[2L_{SBM} + 11^T\right] = 2\,\mathbf{E}\,L_{SBM} + 11^T = 2\,\mathbf{E}\,D_{\mathcal{G}}^{+} - 2\,\mathbf{E}\,D_{\mathcal{G}}^{-} - 2\,\mathbf{E}\,A + 11^T,$$

and $\mathbf{E}\,D_{\mathcal{G}}^{+} = \frac{n}{2}\frac{\alpha\log(n)}{n}$, $\mathbf{E}\,D_{\mathcal{G}}^{-} = \frac{n}{2}\frac{\beta\log(n)}{n}$, and $\mathbf{E}\,A$ is a matrix such with 4 $\frac{n}{2} \times \frac{n}{2}$ blocks where the diagonal blocks have $\frac{\alpha\log(n)}{n}$ and the non-diagonal have $\frac{\beta\log(n)}{n}$. We can write this as $\mathbf{E}\,A = \frac{1}{2}\left(\frac{\alpha\log(n)}{n} + \frac{\beta\log(n)}{n}\right)11^T + \frac{1}{2}\left(\frac{\alpha\log(n)}{n} - \frac{\beta\log(n)}{n}\right)gg^T$

This means that

$$\mathbf{E}\left[2L_{SBM} + 11^T\right] = \left(\left(\alpha-\beta\right)\log n\right)I + \left(1 - (\alpha+\beta)\frac{\log n}{n}\right)11^T - (\alpha-\beta)\frac{\log n}{n}gg^T.$$

Since $2L_{SBM}g = 0$ we can ignore what happens in the span of g and it is not hard to see that

$$\lambda_2\left[\left(\left(\alpha-\beta\right)\log n\right)I + \left(1 - (\alpha+\beta)\frac{\log n}{n}\right)11^T - (\alpha-\beta)\frac{\log n}{n}gg^T\right] = (\alpha-\beta)\log n.$$

This means that it is enough to show that

$$\left\|L_{SBM} - \mathbf{E}\left[L_{SBM}\right]\right\| < \frac{\alpha-\beta}{2}\log n, \tag{4.3}$$

which is a large deviations inequality. ($\|\cdot\|$ denotes operator norm)

I will not describe the details here but the idea is to write $L_{SBM} - \mathbf{E}\left[L_{SBM}\right]$ as a sum of independent random matrices and use Matrix Bernestein in [3].

In fact, using that strategy one can show that, with high probability, 4.3 holds as long as

$$(\alpha-\beta)^2 > 8(\alpha+\beta) + 8/3(\alpha-\beta). \tag{4.4}$$

4.5. **Comparison with phase transition.** To compare (4.4) with (2.4) we note that the latter can be rewritten as

$$(\alpha-\beta)^2 > 4(\alpha+\beta) - 4 \text{ and } \alpha+\beta > 2$$

and so the relaxation achieves exact recovery almost at the threshold, essentially only suboptimal by a factor of 2.

References

[1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *Available online at arXiv:1405.3267 [cs.SI]*, 2014.

[2] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefine programming. *Journal of the Association for Computing Machinery*, 42, 1995.

[3] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv:1004.4389 [math.PR]*, 2010.

[4] L. Vanderberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA
E-mail address: ajsb@math.princeton.edu