MATH888: High-dimensional probability and statistics Fall 2021

Lecture 39 — December 8, 2021

Sebastien Roch, UW-Madison *Scribe: Joowon Lee, Sebastien Roch*

1 Overview

Community recovery remains a popular research area in statistical network analysis, where we seek to find a latent community structure in a network. This technique has been applied to many real-world networks from various scientific domains such as social, biological, physical, and economic contexts. A statistical guarantee should necessarily assume some underlying random graph model. In this lecture, we introduce community recovery under the framework of Stochastic Block Model (SBM), which is a multi-community generalization of the well-known Erdös-Renyí random graph. Specifically, we examined conditions for recovery by conveying relevant notions and investigated a desired estimator for recovery.

This lecture is based on [Abb18].

2 The Stochastic Block Model

Let us consider a simple case with two (strictly) balanced communities.

2.1 Definition

We consider a random graph model on n (even) nodes where there are two communities, say, $+1$ and -1, each consisting of $n/2$ nodes. For two nodes (i, j), we randomly assign an edge between them with probability q_{in} if they belong to the same community, and with probability q_{out} otherwise. Then, the following 2×2 matrix describes the edge density within and across the two communities:

$$W = \begin{array}{c} \\ +1 \\ -1 \end{array} \begin{array}{c} \begin{array}{cc} +1 & -1 \end{array} \\ \begin{bmatrix} q_{in} & q_{out} \\ q_{out} & q_{in} \end{bmatrix} \end{array}.$$

Since nodes belonging to the same community should be more likely to share an edge (at least in some applications), we assume $q_{in} \geq q_{out}$. Let $\mathrm{SBM}(n, q_{in}, q_{out})$ denote the resulting random graph model, which is a probability distribution on the set of all n-node simple graphs. Namely, the model defines an n-vertex random graph with vertices split in two communities, where each vertex is assigned a community label in $\{1, -1\}$ independently under the community prior (q_{in}, q_{out}), and pairs of vertices with labels i and j connect independently with probability $W_{i,j}$ where $i, j \in \{+1, -1\}$.

More specifically, we say that $(X, G) \sim \mathrm{SBM}(n, q_{in}, q_{out})$ if

1. *[Community assignment]* X is uniformly random over

$$\Pi_2^n := \{\mathbf{x} \in \{+1, -1\}^n : \mathbf{x}^T \mathbf{1} = 0\} \text{ where } \mathbf{1} = (1, \cdots, 1)^T$$

which indicates the number of $+1$ and -1 cases is the same *(balanced)*.

2. *[Graph]* G has independent edges where (i, j) is present with probability W_{X_i, X_j} for $\forall i \neq j$.

2.2 Recovery Requirement

The goal of community detection is to estimate (recover) the labels X by observing G. We define the notions of agreement.

Definition 1. *(Agreement) The agreement between two community vectors* $\mathbf{x}, \mathbf{y} \in \{+1, -1\}^n$ *is obtained by maximizing the common components between* $\mathbf{x}$ *and any relabelling of* $\mathbf{y}$*, i.e.,*

$$A(\mathbf{x}, \mathbf{y}) = \max_{s \in \{+1, -1\}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = s(y_i)\}$$

Now consider the following recovery requirements, which is going to be asymptotic, taking place with high probability as n tends to infinity.

Definition 2. *Let* $(X, G) \sim SBM(n, q_{in}, q_{out})$ *and for* $\hat{X} = \hat{X}(G) \in \Pi_2^n$*, an estimate of* X*, we say that we achieve*

- ***Exact recovery:*** $\mathbb{P}(A(X, \hat{X}) = 1) = 1 - o(1)$
- ***Almost exact recovery:*** $\mathbb{P}(A(X, \hat{X}) = 1 - o(1)) = 1 - o(1)$

Then when is it possible to achieve (almost) exact recovery? The following theorem provides the conditions of q_{in} and q_{out} for exact recovery.

Theorem 3. *Exact recovery in* $SBM(n, \alpha \log(n)/n, \beta \log(n)/n)$ *is achievable and efficiently so if and only if* $\sqrt{\alpha} - \sqrt{\beta} > 2$ *and not achievable if* $\sqrt{\alpha} - \sqrt{\beta} < 2$.

2.3 MAP estimator

A natural starting point is to resolve the estimation of X from the noisy observation G by taking the **Maximum A Posteriori (MAP)** estimator.

Let $\Omega(X)$ be the partition corresponding to X and $\hat{\Omega}(G)$ be the partition corresponding to $\hat{X}(G)$. The probability of error (not recovering the true partition), $\mathbb{P}_e$, is given by

$$\mathbb{P}_e := \mathbb{P}(\Omega \neq \hat{\Omega}(G)) = \sum_g \mathbb{P}(\hat{\Omega}(g) \neq \Omega | G = g) \mathbb{P}(G = g),$$

and an estimator $\hat{\Omega}^{MAP}(\cdot)$ minimizing the above must minimize $\mathbb{P}(\hat{\Omega}(g) \neq \Omega | G = g)$ for every g. To minimize $\mathbb{P}(\hat{\Omega}(g) \neq \Omega | G = g)$, we need to choose ω that maximizes the posterior probability

$$
\begin{aligned}
\mathbb{P}(\Omega = \omega | G = g) &= \frac{\mathbb{P}(G = g | \Omega = \omega)\mathbb{P}(\Omega = \omega)}{\mathbb{P}(G = g)} \qquad (\because \text{Bayes rule}) \\
&\propto \mathbb{P}(G = g | \Omega = \omega)\mathbb{P}(\Omega = \omega) \\
&\propto \mathbb{P}(G = g | \Omega = \omega) \qquad \left(\because \mathbb{P}(G = g | \Omega = \omega) = \frac{1}{\#\text{ of partitions}}\right)
\end{aligned}
$$

Then MAP is thus equivalent to the Maximum Likelihood estimator: maximize $\mathbb{P}(G = g | \Omega = \omega)$ over equal size partitions ω.

For fixed g, let $N := N(g)$ be the number of edges in g. For any ω, denote $N_{in} := N_{in}(g, \omega)$ and $N_{out} := N_{out}(g, \omega)$ by the number of edges within and across communities, respectively, and note that $N_{in} = N - N_{out}$. Then

$$
\begin{aligned}
\mathbb{P}(G = g | \Omega = \omega) &= (q_{out})^{N_{out}}(1 - q_{out})^{(\frac{n}{2})^2 - N_{out}}(q_{in})^{N - N_{out}}(1 - q_{in})^{\{\binom{n}{2} - (\frac{n}{2})^2\} - \{N - N_{out}\}} \\
&\propto \left[\frac{q_{out}}{1 - q_{out}} \times \frac{1 - q_{in}}{q_{in}}\right]^{N_{out}}
\end{aligned}
$$

Since we assume $q_{in} \geq q_{out}$, we have $\left[\frac{q_{out}}{1-q_{out}} \times \frac{1-q_{in}}{q_{in}}\right] \leq 1$. Therefore, to maximize $\mathbb{P}(G = g | \Omega = \omega)$, we need to choose ω that minimizes N_{out}. In this sense, MAP is equivalent to solving the *min-bisection problem.*

Alternatively, the same problem can be written as follows:

$$
\max_{\mathbf{x} \in \{+1, -1\}^n, \mathbf{x}^T \mathbf{1} = 0} \mathbf{x}^T A \mathbf{x}
$$

where A is $n \times n$ adjacency matrix. Due to the constraint of $\mathbf{x}^T \mathbf{1} = 0$, it is reasonable to take the second largest eigenvector of A for an appropriate relaxation of MAP. This will be discussed this in more detail next time.

References

[Abb18] Emmanuel Abbe. Community Detection and Stochastic Block Models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, June 2018. Publisher: Now Publishers, Inc.