Recovering Hidden Communities in the Stochastic Block Model

Liam Hardiman

Abstract

In this paper we examine the problem of recovering communities from a random graph $G_{n,p,q}$, produced according to the stochastic block model. In particular, we show that if $p \neq q$ are fixed constants, then we can correctly identify the communities up to a constant number of misclassified vertices with high probability. Additionally, we show that if $p - q = \frac{C}{\sqrt{n}}$ for some constant C, then we can correcctly classify a constant proportion of the vertices.

0.1 Introduction

Suppose there are two schools across the street from one another. Both schools participate in each other's bands, theater productions, and other after-school clubs, so the students interact with each other a fair amount. Say you knew each student's Facebook friend list but not what school they go to. Can you come up with a way separate the students by school from their friends lists alone? We can translate this problem into the language of graph theory.

Divide n vertices into two sets ("communities") of size $n/2$ each. Construct a random graph G by connecting every pair of vertices independently with probability p if they belong to the same community and q if they belong to different communities. The resulting distribution on the set of n-vertex graphs is called the *stochastic block model* and is denoted $G_{n,p,q}$.

As simple examples, let us consider the cases where $p = q$. If $p = q = 0$, then $G_{n,p,q}$ is an empty graph, deterministically. On the other hand, if $p = q = 1$, then we deterministically have the complete graph on n vertices. In the slightly more interesting case where $p = q$ is neither 0 nor 1, we have the Erdős-Rényi random graph $G_{n,p}$ – an interesting object in its own right. If $p = 0$ and $0 < q < 1$, then we have a random bipartite graph.

Is this really any different from the Erdős-Rényi model? That is, does there exist some p' such that the Erdős-Rényi random graph $G_{n,p'}$ behaves just like a random graph using the stochastic block model, $G_{n,p,q}$? If $p' = \frac{n-2}{n-1}\frac{p}{2} + \frac{n}{n-1}\frac{q}{2}$, roughly the average of p and q, then the expected degree of each vertex in $G_{n,p,q}$ and $G_{n,p'}$ is the same. Does this keep us from telling these distributions apart?

Consider the case where $p = \frac{1}{2}$ and $q = \frac{1}{3}$, so $p' = \frac{5}{12}$. Qualitatively speaking, what should $G_{n,p}$ and $G_{n,p,q}$ look like? Using standard concentration inequalities, we can show that each vertex in the $G_{n,p'}$ graph should look roughly the same, i.e. each vertex has degree close to its expected degree. On the other hand, we expect to see two clusters in the $G_{n,p,q}$ graph, with more edges inside the clusters than between them. This clustering behavior, visualized in Figure 1 is even more pronounced when p and q are farther apart, so we can distinguish $G_{n,p'}$ from $G_{n,p,q}$.

In this paper we investigate how to identify the underlying communities in a realization of $G_{n,p,q}$. Of course, if we just randomly guess which community each vertex belongs to, then we expect to correctly classify around half of the vertices with high probability. If p and q are fixed constants, $p \neq q$, it turns out we can do much better.

Theorem 1. *Let $p \neq q$. Then with high probability we can correctly classify all but a constant number of vertices in a realization of $G_{n,p,q}$.*

With a little bit of linear algebra, we address the case where p and q are allowed to depend on n.

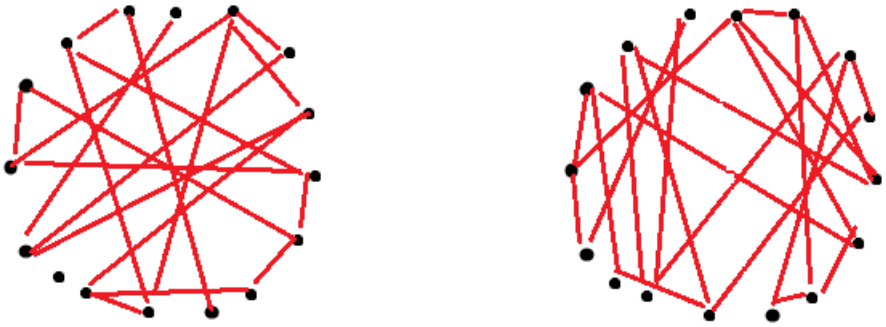

Figure 1: On the left we have $G_{16,1/2}$. On the right we have $G_{16,1/2,1/3}$. Definitely made with Matlab and not MS Paint.

Theorem 2. *Let $0 \leq p, q \leq 1$ be such that $p - q = \frac{100}{\sqrt{n}}$. Then with high probability, we can correctly classify all but $\frac{288p}{(p-q)^2}$ vertices in a realization of $G_{n,p,q}$.*

Our main tool for proving this result is a theorem due to Davis and Kahan.

Theorem 3. *Let B and M be symmetric matrices. Let $R = M - B$. Let α_i be the eigenvalues of B with eigenvectors v_i and μ_i of M with corresponding eigenvectors w_i. Let θ_i be the angle between v_i and w_i. Then*

$$\sin\theta_i \leq \frac{2\|R\|}{\min_{j\neq i}|\mu_i - \mu_j|}.$$

The rest of the paper is organized as follows...