The Geometry of Community Detection

Vaishakhi Mayya[1], Heather Mathews[2], Ricardo Batista[2], Jingwen Zhang[1], Alexander Volfovsky[2], Galen Reeves[1,2]

[1]Department of ECE, Duke University, [2]Department of Statistical Science, Duke University

Overview

The problem of community detection is to partition a network into clusters of nodes (communities) with similar connection patterns. Specific examples include finding like-minded people in a social network and discovering the hierarchical relationships in organizations from observed behavior.

A major limitation of the current analysis of community detection is that it is relevant only to networks exhibiting high levels of homogeneity or symmetry. While the theory provides initial guidelines for how much data one needs to collect, it fails to describe the performance one expects to see in practice. Particularly in settings where individuals belong to multiple communities, there is high variability in the size of the communities, and there is additional covariate information.

The contribution of this work is to study a much broader class of network models in which there can be high variability in the sizes and behaviors of the different communities. Our analysis shows that the performance in these models can be described in terms of a matrix of the effective signal-to-noise ratios (SNRs) that provides a geometrical representation of relationships between the communities. This analysis motivates new methodology for a variety of state-of-the-art algorithms, including spectral clustering, belief propagation, and approximate message passing.

Stochastic Block Model (SBM)

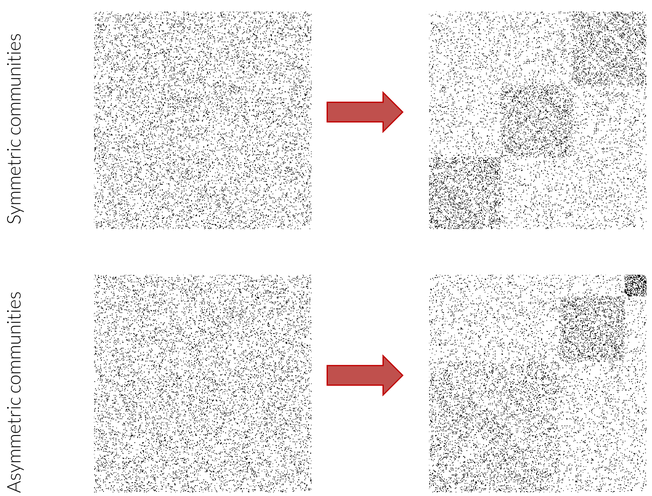
- The SBM is a probabilistic model for a network with n nodes, each of which belongs to one of k communities [Holland & Leinhardt 1983].
- The community label of node i is denoted by binary vector $X_i \in \{0,1\}^k$ with one nonzero entry

$$X_i \overset{iid}{\sim} P = (p_1, \cdots, p_k).$$

- The network is represented by an adjacency matrix $G \in \{0,1\}^{n\times n}$ with $G_{ij} = 1$ if there is an edge between nodes i and j and $G_{ij} = 0$ otherwise. The conditional probability of edges is described by affinity matrix $Q \in [0,1]^{k\times k}$.

$$\Pr\{G_{ij} = 1 | X_1, \dots X_n\} = X_i^T Q X_j.$$

- The degree of a node is the number of edges connected to the node. An SBM is *degree balanced* if the expected degree of a node is independent of its community label.
- The goal of community detection is to recover the labels $\underline{X} = [X_1, \dots, X_n]$ from the observed network G. The SBM parameters (P, Q) are often unknown and also need to be estimated.



Motivating Questions

- What does the geometry of the community structure reveal about the success of recovering community memberships?
- Is there a significant performance gap between existing approaches (e.g., spectral clustering, message passing, or semi-definite programming) and optimal but computationally intractable methods requiring brute-force search?

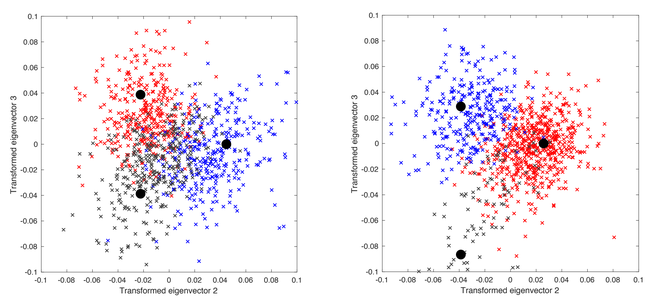
Community Detection

Community detection in a graph with adjacency matrix G, can be studied from two perspectives:

- **Geometry of the eigenvectors**: The eigen-decomposition of an adjacency matrix, G, is often used to identify the community structure of the network. We study the geometry of the eigenvectors and its relation to community detection.
- **Information-theoretic analysis of the SBM**: We can provide a theoretical bound to the mean squared error (MSE) in estimating community membership of the nodes. The bound is obtained by studying the relationship between the problem of community detection and the relatively simpler 'signal-plus-noise problem.'

Eigen-decomposition of Adjacency Matrix

- The eigenvalue spectrum of G typically consists of a dense bulk of closely spaced eigenvalues, plus k outlying eigenvalues separated from the bulk by a significant gap
- The k eigenvectors corresponding to these outliers contain information about the large-scale structure of the network: the largest eigenvector sorts vertices according to their degree, while the remaining $(k-1)$ outlying eigenvectors are correlated with the communities
- When degree distribution is uninformative, spectral methods for community detection cluster on V, the $n \times (k-1)$ matrix composed of the 2^{nd} to k^{th} largest eigenvectors
- Plotting V or, more commonly, a transformed version of V can reveal separable point clusters (i.e., communities), as shown in Figures 1a, 1b. Each plot is a sample of a three-community SBM with respective (Q, P). The black dots represent the expected cluster centers, namely $\mathbb{E}(G|X) = X^T Q X$.



(a) $P = (1/3, 1/3, 1/3)$ (b) $P = (0.6, 0.3, 0.1)$

Figure 1: Examples of transformed eigenvectors V_t for two networks

- In Figures 1a, 1b we plot the transformed version (i.e., V_t) of V, defined as

$$V_t = P + \sqrt{n} B V \text{ where } BB^T = \text{diag}(P) - PP^T. \quad (1)$$

Notice the resemblance between V_t and a Gaussian mixture; this transformation enables inference of the matrix SNR S under the signal-plus-noise framework

Analysis via Connection with Signal-Plus-Noise Problem

- Previous work focusing on the special cases of symmetric SBMs [Deshpande et. al. 2015] and two-community degree balanced SBMs [Lelarge and Miolane 2017] has shown that the information-theoretic limits (i.e., the performance of optimal but possibly intractable methods) can be characterized analytically as a function of the SBM parameters. The formulas are described in terms of a low-dimensional `signal-plus-noise problem' of the form

$$Y = \sqrt{s} X + Z \quad (2)$$

where $X \sim P$ is a k-dimensional binary vector, $Z \sim \mathcal{N}(0, I)$ is standard Gaussian noise, and s is parameter that quantifies the signal-to-noise ratio.
- Our analysis shows that a similar approach can be applied to a much broader class of SBMs with heterogeneous community sizes and connection behaviors. The key innovation is that the scalar signal-to-noise ratio in the signal-plus-noise problem is replaced by a $k \times k$ positive definite matrix S, which is referred to as the matrix SNR [Reeves et. al. 2018],

$$Y = S^{1/2} X + Z. \quad (3)$$

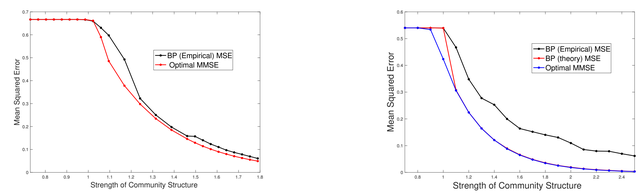
Given SBM parameters (P, Q) the corresponding matrix S is found by solving a optimization problem that requires numerical evaluation of the mutual information $I(X;Y)$.
- The matrix S provides a geometric representation of types of communities structures that can (and cannot) be recovered. For example, if the difference between two entries of X lies in the nullspace of S, then the corresponding communities cannot be differentiated from each other.

Theoretical Threshold of Community Detection Algorithms

- Performance is assessed in terms of mean-squared error of an estimator of the $k \times n$ matrix of community labels. The minimum over all possible estimators is called the minimum mean-squared error (MMSE) and is given by

$$\text{MMSE}(X \mid G) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|X_i - \mathbb{E}[X_i \mid G]\|^2 \right].$$

- The theoretical analysis provides formulas for the asymptotic MMSE and the asymptotic MSE of belief propagation (BP) associated with a sequence of models of increasing size. These formulas are compared with empirical results on a network of size $n = 1000$.



(a) MSE for symmetric SBM with 3 communities and average degree $d = 30$.

(b) MSE for balanced SBM with $P = (0.6, 0.3, 0.1)$ and average degree $d = 30$.

Figure 2: A comparison of the MMSE for degree balanced SBMs. The empirical MSE is computed using the BP algorithm to estimate community memberships. BP is run over 100 sample graphs drawn from an SBM with parameters (P, Q). The x-axis is a measure of the strength of community structure.

Conclusion

Our work bridges and extends recent developments in statistics and information theory to provide theoretical guarantees (MMSE) for general SBMs (e.g., asymmetric, mixed membership).