Provable Overlapping Community Detection in Weighted Graphs

Jimit Majmudar, Stephen Vavasis
University of Waterloo

UNIVERSITY OF WATERLOO

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Community detection (aka graph clustering) is a common unsupervised learning problem which finds applications in diverse domains such as computational biology, social network analysis, and document classification.

Stochastic Blockmodel (SBM) is a popular generative model for random graphs used to perform theoretical analyses of community detection algorithms. However, SBM assumes that the communities do not overlap. Therefore, we use a generalization of SBM, called Mixed Membership Stochastic Blockmodel (MMSB), which allows overlapping communities.

Existing community detection algorithms take as input an unweighted graph, and provide recovery guarantees if the graph is generated according to MMSB. These algorithms are either inefficient, involve multiple tuning parameters whose ideal values cannot be known a priori, are not straightforward to implement, rely on non-convex optimization or make the unrealistic theoretical assumption that each community has node that belongs exclusively that community (such nodes are called pure nodes).

In this work, we provide a simple, efficient, convex-optimization based, provable community detection algorithm which takes as input a weighted graph, and provide recovery guarantees if the graph is generated according to MMSB. Our algorithm involves only one tuning parameter (which is fairly standard, in that it also appears in most other competing algorithms), and our theoretical analysis dispenses the requirement of pure nodes.

MMSB

Suppose n is the number of nodes, k is the number of communities, B is a $k \times k$ matrix.

- Generate an $n \times k$ matrix Θ whose rows are sampled from the Dirichlet distribution.
- Form an $n \times n$ weighted adjacency matrix $P = \Theta B\, \Theta^{\mathrm{T}}$.

Question: Given P, how to efficiently recover a good approximation of Θ?

Algorithm

Algorithm 1 SP+LP

Input: Matrix P generated according to MMSB, number of communities k
Output: Estimated characteristic vectors $\hat{\theta}_1, \ldots, \hat{\theta}_k \in [0,1]^n$

1: $\mathcal{J} = \text{SuccessiveProjection}(P)$
2: **for** $i \in [k]$ **do**
3: $\quad (\mathbf{x}^*, \mathbf{y}^*) = \arg\min_{(\mathbf{x},\mathbf{y})} \mathbf{e}^T\mathbf{x}$ s.t. $\mathbf{x} \geq \mathbf{0}, x_{\mathcal{J}(i)} \geq 1, \mathbf{x} = P\mathbf{y}$
4: $\quad \hat{\theta}_i = \mathbf{x}^*/\|\mathbf{x}^*\|_\infty$
5: **end for**

Interpretation: first determine k *almost pure nodes*, i.e. nodes which almost exclusively belong to one community, using SuccessiveProjection subroutine; then use them to determine the columns of Θ using k linear programs, which can be interpreted as obtaining a sparse, non-negative basis for the column range of P.

Main Theoretical Result

Theorem 3.1. *Suppose $k \geq 2$, B is full-rank, and all k parameters of the Dirichlet distribution are equal to $\alpha \in \mathbb{R}$. Let $w := 8\kappa\sqrt{\alpha k + 1}$ and define*

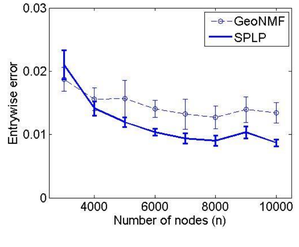
$$\epsilon_1 := \min\left(\frac{1}{\sqrt{k-1}}, \frac{1}{2}\right) \frac{1}{2\sqrt{2}w(1+80w^2)}$$
$$\epsilon_2 := \frac{7}{3520\sqrt{2}kw^2}.$$

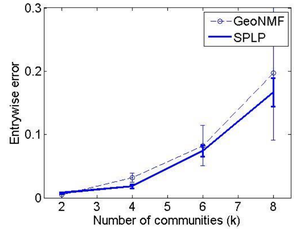
If $n > \dfrac{\log(p/k)}{\log I_{1-\epsilon}(\alpha, (k-1)\alpha)}$ for some $p \in (0,1)$ and $\epsilon \in (0, \min\{\epsilon_1, \epsilon_2\})$, then there exists a permutation π of the set $[k]$ such that vectors $\hat{\theta}_1, \ldots, \hat{\theta}_k$ returned by SP+LP satisfy

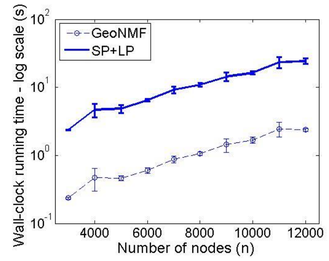
$$\max_{j \in [k]} \|\hat{\theta}_j - \theta_{\pi(j)}\|_\infty = \mathcal{O}(\alpha k^2 \kappa^2 \epsilon) \tag{3}$$

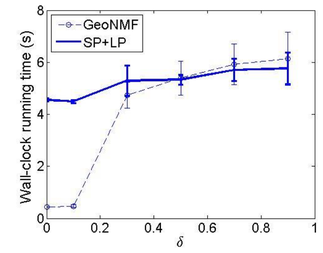
with probability at least $1 - p - c_1 e^{-c_2 n}$ where c_1, c_2 are constants that depend on α, k, κ.
(Here $I_x(y,z)$ denotes the regularized incomplete beta function.)

Experiments on Synthetic Graphs









For the figure on right, we have $B = (1-\delta)I + \delta J$ where J is the matrix of all ones.

Experiments on Real-World Networks

We perform overlapping community detection on protein-protein interaction (PPI) networks to estimate protein complexes (clusters of functionally-similar proteins).

Table 1: Comparision of SP+LP with ClusterONE on Krogan core, Krogan extended, and Gavin datasets using SGD repository as validation set.

Validation set	Metric	Krogan core		Krogan extended		Collins	
		SP+LP	ClusterONE	SP+LP	ClusterONE	SP+LP	ClusterONE
SGD	MMR	0.389	0.418	0.428	0.364	0.372	0.532
	frac	0.598	0.667	0.632	0.594	0.557	0.828
	GA	0.525	0.663	0.542	0.628	0.504	0.731
	Score	1.512	1.748	1.602	1.586	1.433	2.091

Table 2: Comparision of SP+LP with ClusterONE on Krogan core, Krogan extended, and Gavin datasets using MIPS repository as validation set.

Validation set	Metric	Krogan core		Krogan extended		Collins	
		SP+LP	ClusterONE	SP+LP	ClusterONE	SP+LP	ClusterONE
MIPS	MMR	0.285	0.317	0.319	0.282	0.275	0.418
	frac	0.537	0.669	0.576	0.573	0.547	0.782
	GA	0.331	0.438	0.336	0.422	0.397	0.555
	Score	1.153	1.424	1.231	1.277	1.219	1.755

Funding Disclosure

Support for this work was provided by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada and by internal University of Waterloo funds.