Community Extraction for Social Networks

Yunpeng Zhao

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

August 3, 2011

Advisor: Liza Levina and Ji Zhu

Outline

- Review of community detection
- Community extraction
- Asymptotic consistency
- Simulation study
- Real data analysis

Network data

Network analysis has been a focus of attention in different fields.

- Social science: friendship networks
- Internet: WWW, hyper-links
- Biology: food webs, gene regulatory networks

Community detection

- Communities: Networks consist of communities, or clusters, with many connections within a community and few connections between communities.
- Community detection problem: For an undirected network $N = (V, E)$, the community detection problem is typically formulated as finding a partition $V = V_1 \cup \cdots \cup V_K$ which gives "tight" communities in some suitable sense.

Community detection problem

Existing community detection methods: minimizing links between communities while maximizing links within communities (see Newman (2004) for a review).

For simplicity, we consider the case of partitioning the network into two communities V_1 and V_2.

Min-cut (Wu and Leahy, 1993)

To minimize

$$R = \sum_{i \in V_1, j \in V_2} A_{ij} \ .$$

However, min-cut always yields a trivial solution of $V_1 = V$ or $V_2 = V$.

Ratio-cut (Wei and Cheng, 1989)

$$\min R/(|V_1| \cdot |V_2|),$$

where $|V_1|$ and $|V_2|$ represent the sizes of two groups respectively.

Ratio-cut can avoid trivial solutions because the maximizer of $|V_1| \cdot |V_2|$ is achieved at $|V_1| = |V_2| = |V|/2$.

Normalized-cut (Shi and Malik, 2000)

$$\min \frac{R}{\operatorname{assoc}(V_1, V)} + \frac{R}{\operatorname{assoc}(V_2, V)},$$

where $\operatorname{assoc}(V_k, V) = \sum_{i \in V_k, j \in V} A_{ij}$ for $k = 1, 2$.

Normalized-cut can avoid trivial solutions because an extremely small group V_k may have a large ratio $R / \operatorname{assoc}(V_k, V)$.

Modularity (Newman and Girvan, 2004)

To maximize

$$Q = \sum_{k=1}^{2} \left[\frac{O_{kk}}{L} - \left(\frac{D_k}{L} \right)^2 \right],$$

where $O_{kk} = \sum_{i \in V_k, j \in V_k} A_{ij}, D_k = \sum_{i \in V_k, j \in V} A_{ij}, L = \sum_{k=1}^{2} D_k$.

Q represents the fraction of edges that fall within communities, minus the "average" value of the same quantity if edges fall at random given the degree of each node.

Outline

- ✓ Review of community detection
- Community extraction
- Asymptotic consistency
- Simulation study
- Real data analysis

Community extraction

- Most networks consist of a number (not known a priori) of communities, with relatively tight links within each community and sparse links to the outside, and “background” nodes that only have sparse links to other nodes.
- We propose a method that extracts communities sequentially: at each step, the tightest is extracted from the network until no more meaningful communities exist.

Criterion

- Extract one community at a time by looking for a set of nodes with a large number of links within itself and a small number of links to the rest of the network.
- The links within the complement of this set do not matter.

To maximize

$$W(S) = \frac{I(S)}{k^2} - \frac{B(S)}{k(n-k)} ,$$

where

$$I(S) = \sum_{i,j \in S} A_{ij} , \; B(S) = \sum_{i \in S, j \in S^c} A_{ij} , \; k = |S| .$$

Adjusted criterion

- Empirically, the previous criterion performs well for dense networks. However, it always finds very small communities for sparse networks.
- To avoid small communities, we also propose

To maximize

$$W_a(S) = k(n-k)\left(\frac{I(S)}{k^2} - \frac{B(S)}{k(n-k)}\right).$$

The factor $k(n-k)$ penalizes communities with k close to 1 or n and encourages more balanced solutions.

Algorithm

- Tabu Search (Glover, 1986; Glover and Laguna, 1997): a local optimization technique based on label switching
- Run the algorithm for many randomly ordered nodes

Outline

Block models

Asymptotic consistency can be established under the assumption of block models.

General block models

1. Each node is assigned to a block independently of other nodes, with probability π_k for block k, $1 \leq k \leq K$, $\sum_{k=1}^{K} \pi_k = 1$.
2. Given that node i belongs to block a and node j belongs to block b, $P[A_{ij} = 1] = p_{ab}$, and all edges are independent.

Block models for networks with background

- We can define the last block as background, by assuming $p_{aK} < p_{bb}$ for all $a = 1, \ldots, K$, and all $b = 1, \ldots, K-1$.

Asymptotic consistency

- For simplicity, assume there is only one community and background in the network ($K = 2$ with parameters $p_{11}, p_{12}, p_{22}, \pi$ and $1 - \pi$).
- Let $\boldsymbol{c}$ denote the true community labels, $\hat{\boldsymbol{c}}^{(n)}$ denote the estimated labels, based on Bickel and Chen (2010), we proved

Theorem

For any $0 < \pi < 1$, if $p_{11} > p_{12}$, $p_{11} > p_{22}$ and $p_{11} + p_{22} > 2p_{12}$, the maximizer $\hat{\boldsymbol{c}}^{(n)}$ of both unadjusted and adjusted criteria satisfies

$$P[\hat{\boldsymbol{c}}^{(n)} = \boldsymbol{c}] \rightarrow 1 \quad as \quad n \rightarrow \infty.$$

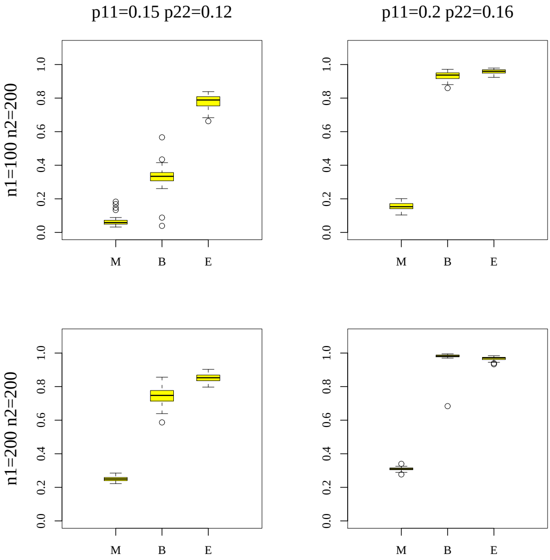
Outline

- ✓ Review of community detection
- ✓ Community extraction
- ✓ Asymptotic consistency
- Simulation study
- Real data analysis

Simulation I

- Two communities with background (block model)
- $n = 1000$
- $n_1 = 100, 200, n_2 = 100$
- $p_{12} = p_{23} = p_{13} = p_{33} = 0.05$
- $p_{11} = 0.05i, p_{22} = 0.04i, i = 3, 4$
- Rand index

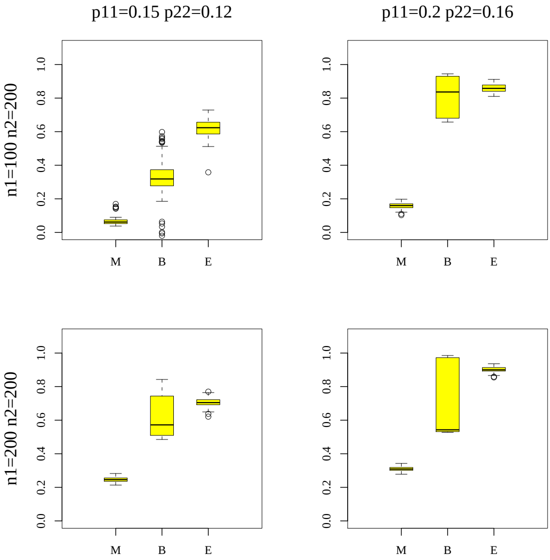
Results for simulation I



Simulation II

- Two communities with background
- $n = 1000$
- $n_1 = 100, 200, n_2 = 100$
- $p_{12} = p_{23} = p_{13} = p_{33} = 0.05$
- $p_{11} = 0.05i, p_{22} = 0.04i, i = 3, 4$
- Doubling the degree for 10 highest degree nodes

Results for simulation II



Outline

- ✓ Review of community detection
- ✓ Community extraction
- ✓ Asymptotic consistency
- ✓ Simulation study
- Real data analysis

Karate club network

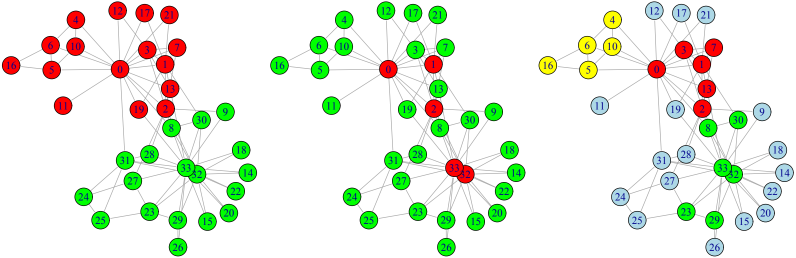
- Friendships between 34 members of a karate club (Zachary, 1977).
- This club has subsequently split into two parts following a disagreement between an instructor (node 0) and an administrator (node 33).

Karate club network

(a) Modularity

(b) Block model

(c) Extraction

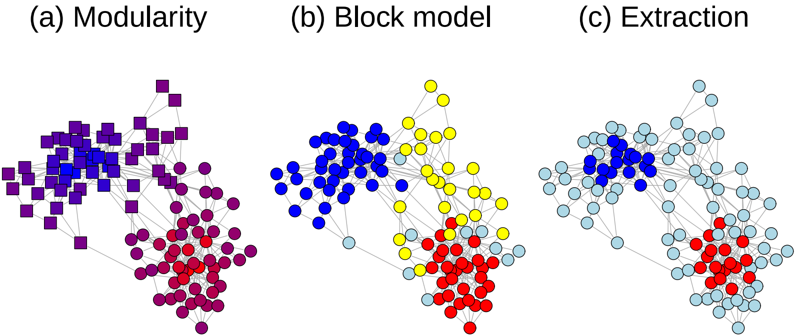


Political books network

Links in the political books network (Newman, 2006) represent pairs of books frequently bought together on amazon.com.

Blue: liberal
Red: conservative

Political books network



Acknowledgment

- Thank my advisors: Elizaveta Levina and Ji Zhu
- Thank Professor Mark Newman for constructive suggestion and sharing his code
- Thank my friends: Xi Chen, Jian Guo and Yizao Wang

And also

Thank you all very much!