Consistency of community detection for networks under degree-corrected block models

Yunpeng Zhao

Department of Statistics, University of Michigan

Joint work with Elizaveta Levina and Ji Zhu

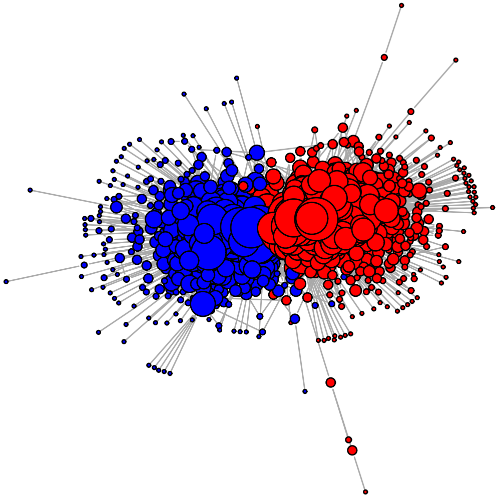
Outline

- Models for community detection
- Consistency of community detection criteria
- Simulation study
- Data example
- Conclusion

What is a network?

A network is a graph $N = (V, E)$, where V is the set of nodes and E is the set of edges.

N may be directed or undirected, weighted or unweighted.



Network data

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

- Social science: friendship networks, collaboration networks
- Computer science: computer networks, internet
- Biology: gene regulatory networks, protein-protein networks

From a statistical point of view

A network is a $n \times n$ random matrix $A = [A_{ij}]$. One may put a probability distribution $\mathbb{P}$ on A.

1. Test goodness of fit
2. Parameter estimation
3. Statistical inference

We only focus on undirected and unweighted networks: A is a symmetric binary random matrix.

Community detection

- Communities: Networks consist of communities, or clusters, with many connections within communities but few connections between communities.
- Community detection problem: For an undirected network $N = (V, E)$, the community detection problem is typically formulated as finding a disjoint partition $V = V_1 \cup \cdots \cup V_K$ with each V_k being a community.

Community detection methods

- Algorithm-based: Hierarchical clustering, edge removal, etc.
- Criterion-based: Ratio cut (Wei & Cheng,1989), normalized cut (Shi & Malik,2000), modularity (Newman,2006), community extraction (Zhao et al.,2011), etc.
- Model-based: Block model (Bickel & Chen,2010), degree-corrected block model (Karrer & Newman,2010), etc.

Block models (Holland, 1983)

1. Each node is assigned with a community label c_i, and the labels c_i are generated independently from $Multinomial(\pi)$ with $\pi = (\pi_1, \ldots, \pi_K)^T$.
2. Given $\boldsymbol{c}$, the edges A_{ij} are independent Bernoulli random variables with $\mathbb{P}(A_{ij} = 1|\boldsymbol{c}) = P_{c_i c_j}$, where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix.

"Null model" ($K = 1$): Erdos-Renyi graph (all edges form independently w.p. p).

Degree-corrected block models (Karrer & Newman, 2010)

1. Each node is assigned with a community label c_i, and the labels c_i are generated independently from $Multinomial(\pi)$ with $\pi = (\pi_1, \ldots, \pi_K)^T$.
2. In addition to community label c_i, each node is associated with a latent variable θ_i, which reflects degree variations, where $\mathbb{E}[\theta_i] = 1$.
3. Given $\boldsymbol{c}$ and θ, the edges A_{ij} are independent Bernoulli random variables with $\mathbb{P}(A_{ij} = 1 | \boldsymbol{c}, \theta) = \theta_i \theta_j P_{c_i c_j}$, where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix.

$\theta_i \equiv 1$ gives the standard block model.

"Null model": the expected degree random graph (all edges form independently with $P(A_{ij} = 1) \propto d_i d_j$).

Notation

- For any community label assignments $\boldsymbol{e} = \{e_1, ..., e_n\}$, define $O(\boldsymbol{e}) = [O_{kl}(\boldsymbol{e})]$, where

$$O_{kl} = \sum_{ij} A_{ij} I\{e_i = k, e_j = l\},$$

$$O_k = \sum_{l} O_{kl},$$

  and $O_k = \sum_l O_{kl}$, $L = \sum_{kl} O_{kl}$, $n_k = \sum_k I\{e_i = k\}$.
- Note $O(\boldsymbol{e})$ does not depend only on true labels $\boldsymbol{c}$.

Likelihood-type criteria

Maximize likelihood of the block model (Bickel & Chen, 2010) :

$$\max_{\boldsymbol{e}} Q_{BL}(\boldsymbol{e}) = \sum_{kl} O_{kl} \log \frac{O_{kl}}{n_k n_l}$$

Maximize likelihood of the degree-corrected block model (Karrer & Newman, 2010):

$$\max_{\boldsymbol{e}} Q_{DCBL}(\boldsymbol{e}) = \sum_{kl} O_{kl} \log \frac{O_{kl}}{O_k O_l}$$

Modularity-type criteria

Maximize the difference between observed number of edges within communities and expected number of edges under the null model:

$$\max_{\boldsymbol{e}} Q(\boldsymbol{e}) = \sum_{ij} [A_{ij} - P_{ij}] I(e_i = e_j),$$

where P_{ij} is the (estimated) probability of an edge falling between i and j under the null model.

Modularity-type criteria

- When the null model is ER graph, $P_{ij} = L/n^2$ and $Q(\boldsymbol{e})$ becomes

$$\max_{\boldsymbol{e}} Q_{ERM}(\boldsymbol{e}) = \sum_k (O_{kk} - \frac{n_k^2}{n^2} L).$$

- When the null model is the expected degree random graph, $P_{ij} = k_i k_j / L$ and $Q(\boldsymbol{e})$ becomes

$$\max_{\boldsymbol{e}} Q_{NGM}(\boldsymbol{e}) = \sum_k (O_{kk} - \frac{O_k^2}{L}).$$

This is the well-known Newman-Girvan Modularity.

Notation

- A fundamental question: consistency – whether a detection method can recover the true community labels.
- For any estimator $\hat{\boldsymbol{c}}$ of $\boldsymbol{c}$, we call $\hat{\boldsymbol{c}}$ is consistent if

$$\mathbb{P}[\hat{\boldsymbol{c}} = \boldsymbol{c}] \to 1.$$

- For simplicity, assume θ_i in the degree-corrected block model is discrete, $\mathbb{P}(c_i = k, \theta_i = d_m) = \Pi_{km}$.
- For any k, define $\tilde{\pi}_k = \sum_m d_m \Pi_{km}$.
- Define $\tilde{Q} = \sum_{kk'} \tilde{\pi}_k \tilde{\pi}_k' P_{kk'}, \tilde{W}_{kk'} = \frac{\tilde{\pi}_k \tilde{\pi}_k' P_{kk'}}{\tilde{Q}}$, and $\tilde{\mathscr{E}} = \tilde{W} - (\tilde{W}\mathbf{1})(\tilde{W}\mathbf{1})^T$.

Consistency of likelihood-type criteria

Theorem

NGM is consistent under the degree-corrected block model with the parameter constraint $\tilde{\mathscr{E}}_{kk} > 0, \tilde{\mathscr{E}}_{kk'} < 0 \quad$ *for all* $k \neq k'$, *When* $K = 2$, *the condition can be simplified as*

$$P_{11}P_{22} > P_{12}^2.$$

Theorem

ERM is consistent under the block model with the parameter constraint $P_{kk} > Q, P_{kk'} < Q \quad$ *for all* $k \neq k'$, *where* $Q = \sum_{kk'} \pi_k \pi_{k'} P_{kk'}$.

Consistency of likelihood-type criteria

Theorem

BL is consistent under the block model.

Theorem

DCBL is consistent under both the block model and the degree-corrected block model.

Summary of community detection criteria

	Without correction	With correction
Modularity-type	$\sum_k (O_{kk} - \frac{n_k^2}{n^2} L)$ (ERM)	$\sum_k (O_{kk} - \frac{O_k^2}{L})$ (NGM)
Likelihood-type	$\sum_{kl} O_{kl} \log \frac{O_{kl}}{n_k n_l}$ (BL)	$\sum_{kl} O_{kl} \log \frac{O_{kl}}{O_k O_l}$ (DCBL)

A general theorem on consistency under degree-corrected block models

Theorem

For any Q that can be written as

$$Q(\boldsymbol{e}) = F\left(\frac{O}{n^2}, \left[\frac{n_1}{n}, \ldots, \frac{n_K}{n}\right]^T\right),$$

under some regularity conditions and the following:

(*) $F(G(R), \sum_{lm} R_{.lm})$ *is uniquely maximized over* $\{R : R \geq 0, \sum_k R_{k..} = \Pi\}$ *by* $R_{klm} = \Pi_{lm}\delta_{kl}$ *for any* m*, where* $G \in \mathscr{R}^{K \times K}, R \in R^{K \times K \times M}$*,* $G(R) = \sum_{ll'mm'} \theta_m \theta_{m'} P_{ll'} R_{klm} R_{k'l'm'}, R_{klm} = \frac{1}{n} \sum_{i=1}^{n} I(e_i = k, c_i = l, \theta_i = d_m)$.

Q is consistent under degree-corrected block models.

(*) says that the “population” version of Q is maximized by the correct assignment.

Simulation study

- We consider networks with 1000 nodes and 2 communities, and the matrix P

$$P = \begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}.$$

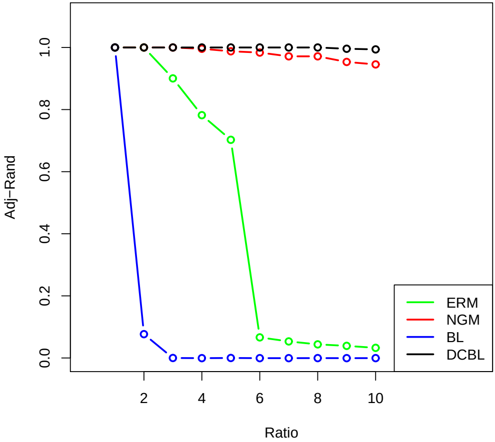
- Adjusted Rand index: a measure of the similarity between two community partitions with 1 being perfect match, and 0 begin the expected agreement between 2 random partitions.

Degree-corrected block model

Fix $\pi_1 = 0.3, \pi_2 = 0.7$.

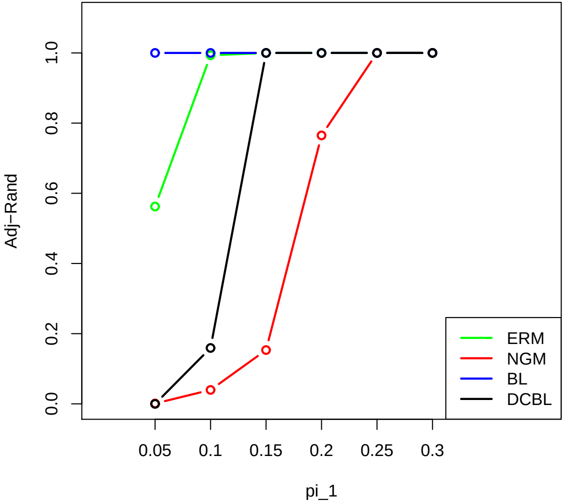
$$\theta = \begin{cases} d_1 & \text{w.p.} \frac{1}{2}, \\ d_2 & \text{w.p.} \frac{1}{2}. \end{cases}$$

The ratio d_1/d_2 changes from 1 to 10.



Block model

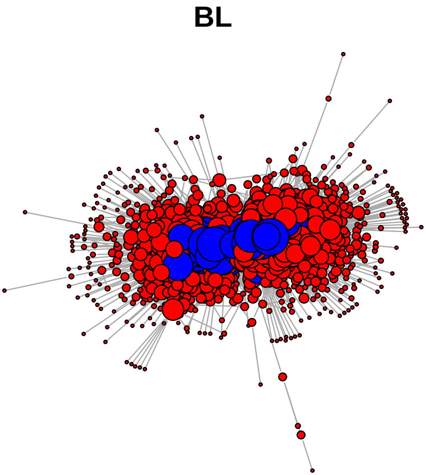
Block model with π_1 changing from 0.05 to 0.3



A network of political blogs

- Nodes: blogs on US politics ($n = 1222$). Edges: hyper-links between blogs. (Adamic&Glance(2005))
- BL, split of high-degree and low-degree nodes.
- DCBL, is very close to the true split of liberal and conservative blogs. ERM and NGM yield very similar results to DCBL.

A network of political blogs



Consistency results

- NGM and DCBL are consistent under the block model with or without degree-correction. But ERM and BL are only consistent under the block model without degree-correction.
- ERM and NGM are consistent with some parameter constraints. But BL and DCBL are consistent for all parameter settings.

Finite sample performance

- BL and DCBL work best where their model assumptions are correct.
- ERM is more robust than NGM under the block model with unbalanced community sizes.
- ERM is more robust than BL under the degree-corrected block model.