On consistency of community detection in networks

Yunpeng Zhao

Department of Statistics, George Mason University

Joint work with: Elizaveta Levina and Ji Zhu

Outline

1. Consistency of community detection criteria under degree-corrected block models

2. Community extraction

Network data

Network data appear in many fields:

- Social and friendship networks, citation networks
- World Wide Web
- Gene regulatory networks, food webs

Definition of networks

A network $N = (V, E)$: V is the set of nodes, $|V| = n$, E is the set of edges

- N is represented by its $n \times n$ adjacency matrix A:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from node } i \text{ to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

- A can be symmetric (undirected networks) or asymmetric (directed networks).
- We only focus on undirected networks.

From a statistical point of view

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

Examples of network models:

- Block models (Holland et al 1983, Faust & Wasserman 1992)
- Exponential Random Graph Models (Robins et al 2006)
- Latent space models (Hoff et al 2002).

Statistical questions

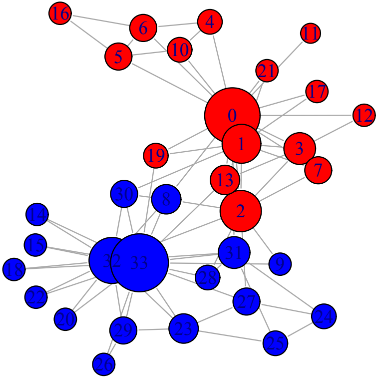
1. Test goodness of fit (Hunter et al 2008)
2. Fitting models (Bickel & Chen 2009, Snijders 2002)
3. Statistical inference and uncertainty assessment (Chatterjee & Diaconis 2011, Shalizi & Rinaldo 2011)

Community detection

- An important topic: community detection
- Communities are cohesive groups of nodes
- Most common interpretation: many links within and few links between
- The community detection problem is typically formulated as finding a disjoint partition $V = V_1 \cup \cdots \cup V_K$

Example: Karate club

A friendship network of a karate club (Zachary 1977), split into two groups, which can be used as “ground truth”.
Node size is proportional to degree.



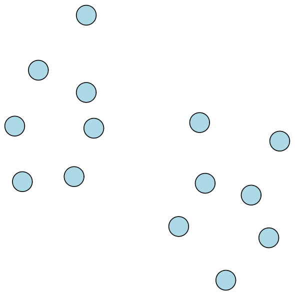
Community detection methods

Existing methods can be loosely classified into three categories.

- Greedy algorithms:
  hierarchical clustering, edge removal (Girvan & Newman 2002)
- Optimizing a global criterion over all partitions:
  normalized cuts (Shi & Malik 2000), modularity (Newman 2006), extraction (Zhao et al 2011b), and many others
- Fitting a model for a network with communities:
  block models (Bickel & Chen 2009), degree-corrected block models (Karrer & Newman 2010), and others

Block model

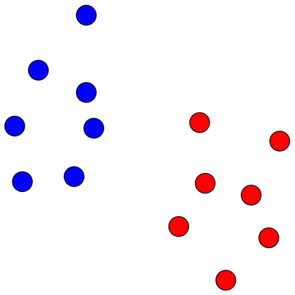
Holland et al (1983)



Block model

Holland et al (1983)

1. Each node is independently assigned a community label c_i, multinomial with parameter $\pi = (\pi_1, \ldots, \pi_K)^T$.



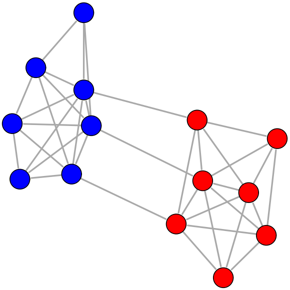
Block model

Holland et al (1983)

1. Each node is independently assigned a community label c_i, multinomial with parameter $\pi = (\pi_1, \ldots, \pi_K)^T$.
2. Given node labels $\boldsymbol{c}$, the edges A_{ij} are independent Bernoulli random variables with

$$P(A_{ij} = 1) = P_{c_i c_j} ,$$

where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix.



Block model

- Fitting: MCMC (Snijders & Nowicki 1997), profile likelihood (Bickel & Chen 2009), or variational approach (Daudin et al 2008)
- The “null” model ($K = 1$): the Erdos-Renyi graph (all edges form independently with probability p)
- Limitation: node degrees within one community are homogeneous, which does not allow for “hubs”–nodes with very high degrees.

Degree-corrected block model

Karrer & Newman (2010)

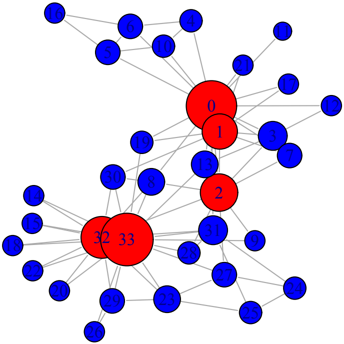
- Generalizes the block model to allow for varying degrees within communities
- Each node is associated with a degree parameter θ_i, and

$$P(A_{ij} = 1) = \theta_i \theta_j P_{c_i c_j} \ .$$

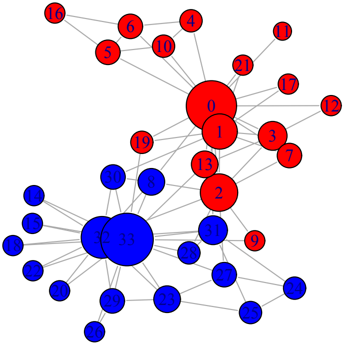
- The standard block model corresponds to $\theta_i \equiv const$.
- The "null" model ($K = 1$): the expected degree random graph, a.k.a. configuration model (all edges form independently with $P(A_{ij} = 1) \propto \theta_i \theta_j$).
- Fits a number of datasets better than the block model

Example: Karate club

Block model



With degree-correction



Notation

For any community label assignment $\boldsymbol{e} = \{e_1, ..., e_n\}$, $e_i \in \{1, \ldots, K\}$, define

$$O_{kl} = \sum_{ij} A_{ij} I\{e_i = k, e_j = l\}, \text{ \# edges between communities } k \text{ and } l$$

$$O_k = \sum_{l} O_{kl}, \text{total degrees in community } k$$

$$L = \sum_{kl} O_{kl}, \text{ total \# edges}$$

$$n_k = \sum_{k} I\{e_i = k\}, \text{ \# nodes in community } k$$

Depend only on the data

Likelihood

Maximize the profile likelihood of the block model (Bickel & Chen 2009) :

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

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

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

Modularity

Maximize observed number of edges within communities minus expected under a null model, over all label assignments $\boldsymbol{e}$:

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

where $E[A_{ij}]$ is the (estimated) expectation under the null model.

Modularity

- When the null model is Erdos-Renyi graph, $E[A_{ij}] = L/n^2$ and $Q(\boldsymbol{e})$ becomes

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

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

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

This is the well-known Newman-Girvan Modularity.

Community detection criteria

	Block model	Degree correction
Modularity	$\sum_k (O_{kk} - \frac{n_k^2}{n^2} L)$	$\sum_k (O_{kk} - \frac{O_k^2}{L^2} L)$
Likelihood	$\sum_{kl} O_{kl} \log \frac{O_{kl}}{n_k n_l}$	$\sum_{kl} O_{kl} \log \frac{O_{kl}}{O_k O_l}$

- The block model measures "community size" by the number of nodes, and the degree-corrected block model by the number of edges.
- Modularity encourages the number of edges within communities larger than the average.

Consistency of label assignments

- Strong consistency (Bickel & Chen 2009): A label estimator $\hat{\boldsymbol{c}}$ is strongly consistent if

$$\mathbb{P}[\hat{\boldsymbol{c}} = \boldsymbol{c}] \to 1, \text{ as } n \to \infty.$$

- Weak consistency: A label estimator $\hat{\boldsymbol{c}}$ is weakly consistent if

$$\forall \varepsilon > 0, \, \mathbb{P}\left[\left(\frac{1}{n}\sum_{i=1}^{n} 1(\hat{c}_i \neq c_i)\right) < \varepsilon\right] \to 1, \text{ as } n \to \infty.$$

Consistency of label assignments

- Parametrize the probability matrix by $P_n = \rho_n P$, where $\rho_n = P(A_{ij} = 1)$ is the probability of an edge, and $\lambda_n = n\rho_n$ is the average expected degree of the graph.
- Strong consistency assumes that $\frac{\lambda_n}{\log n} \to \infty$.
- Weak consistency assumes that $\lambda_n \to \infty$.

A variant of the degree-corrected block model

Our interpretation of Karrer & Newman

- Given node labels $\boldsymbol{c}$, each node is independently assigned a discrete "degree variable" θ_i, with $E[\theta_i] = 1$ for identifiability.
- Given $\boldsymbol{c}$ and θ, the edges A_{ij} are independent Bernoulli random variables with

$$P(A_{ij} = 1 | \boldsymbol{c}, \theta) = \theta_i \theta_j P_{c_i c_j} \ .$$

A general theorem on consistency under degree-corrected block models

Theorem (Zhao, Levina, and Zhu 2011a)

For any criterion Q of the form

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

if F satisfies some regularity conditions and its population version is uniquely maximized by the true partition, then Q is consistent under degree-corrected block models.

Notation

- For simplicity, assume θ_i in the degree-corrected block model is discrete, $\mathbb{P}(c_i = k, \theta_i = d_m) = \Pi_{km}$.

Notation

- 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 $\tilde{\pi}_k = \sum_m d_m \Pi_{km}$. (For the standard block model, $\tilde{\pi}_k = \pi_k$.)
- Define $\tilde{P}_0 = \sum_{kk'} \tilde{\pi}_k \tilde{\pi}_k' P_{kk'}, \widetilde{W}_{kk'} = \frac{\tilde{\pi}_k \tilde{\pi}_k' P_{kk'}}{\tilde{P}_0}$, and $\tilde{\mathscr{E}} = \widetilde{W} - (\widetilde{W}\mathbf{1})(\widetilde{W}\mathbf{1})^T$.

Consistency of modularity

Theorem (Zhao, Levina, and Zhu 2011a)

Newman-Girvan modularity 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

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

Consistency of modularity

Theorem (Zhao, Levina, and Zhu 2011a)

Newman-Girvan modularity 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 (Zhao, Levina, and Zhu 2011a)

Erdos-Renyi modularity is consistent under the block model with the parameter constraint $P_{kk} > P_0, P_{kk'} < P_0 \quad$ for all $k \neq k'$, where $P_0 = \sum_{kk'} \pi_k \pi_{k'} P_{kk'}$.

Consistency of likelihood

Theorem (Bickel & Chen 2009)

Block model likelihood is consistent under the block model.

Consistency of likelihood

Theorem (Bickel & Chen 2009)

Block model likelihood is consistent under the block model.

Theorem (Zhao, Levina, and Zhu 2011a)

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

Summary of consistency results

- Likelihoods are always consistent under their assumed model

Summary of consistency results

- Likelihoods are always consistent under their assumed model
- Modularities are consistent under their assumed model under a parameter constraint indicating stronger links within than between

Summary of consistency results

- Likelihoods are always consistent under their assumed model
- Modularities are consistent under their assumed model under a parameter constraint indicating stronger links within than between
- Anything consistent under degree-corrected block model is also consistent under the block model as a special case

Summary of consistency results

- Likelihoods are always consistent under their assumed model
- Modularities are consistent under their assumed model under a parameter constraint indicating stronger links within than between
- Anything consistent under degree-corrected block model is also consistent under the block model as a special case
- Methods designed under the block model assumption are not generally consistent under the degree-corrected block model

Simulation study

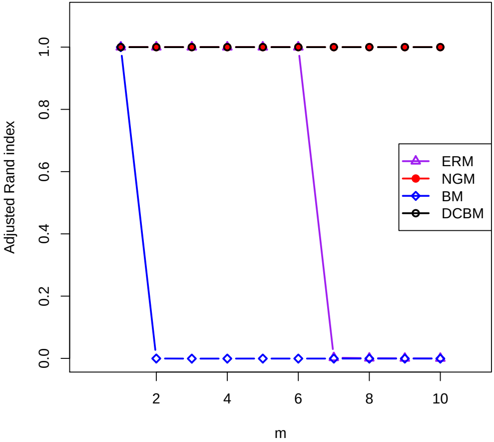
- Let $n = 1000$, $K = 2$, and $P = \begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}$.
- Let θ_i take two values d_1 and d_2 with probability 0.5 each, independently of $\boldsymbol{c}$
- Measure agreement by adjusted Rand index, a measure of similarity between two partitions:
  1 is perfect match;
  0 is expected agreement between two random partitions.

Degree-corrected block model

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

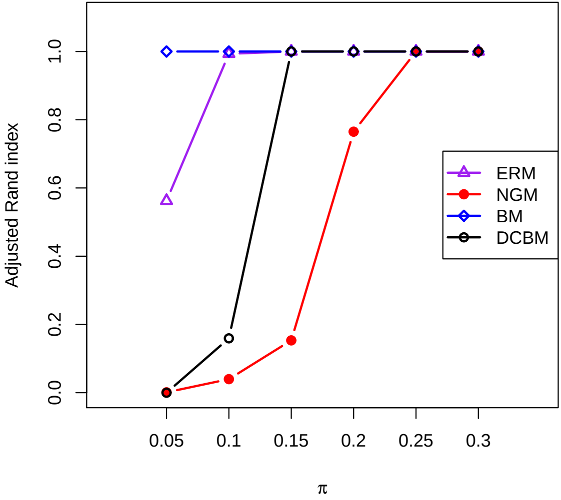
$$\theta = \begin{cases} d_1 & \text{w.p.} \frac{1}{2}, \\ d_2 & \text{w.p.} \frac{1}{2}. \end{cases}$$

The ratio d_1/d_2 changes from 1 to 10.



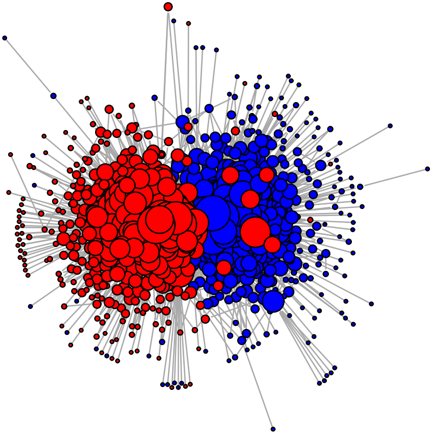
Block model

Block model with π_1 changing from 0.05 to 0.3

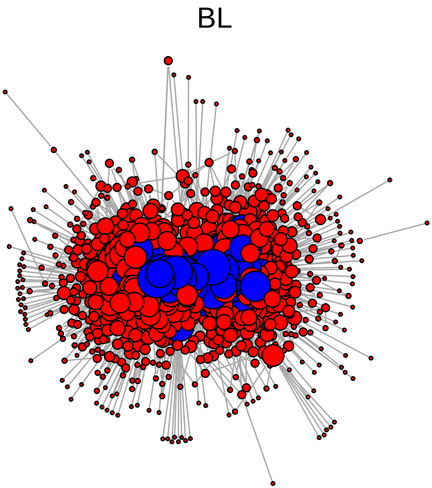


A network of political blogs

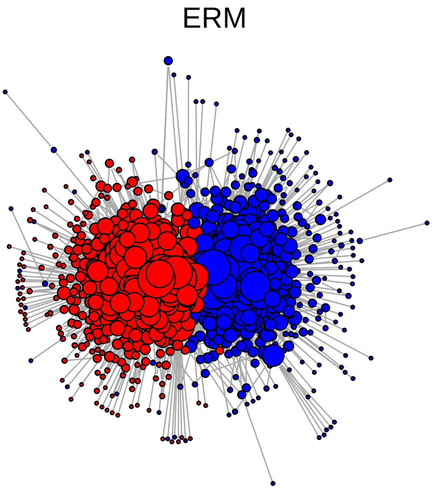
Adamic & Glance (2005) manually labeled 1222 blogs as liberal or conservative, represented by colors, edges are web links (we ignore direction). Node size is proportional to log degree.



A network of political blogs



A network of political blogs



Outline

- ✓ Consistency of community detection criteria under degree-corrected block models

- Community extraction

Limitations of partition methods

- Many real-world networks contain nodes with few links that may not belong to any community (“background”)
- Determining the number of communities in advance is difficult

Community extraction

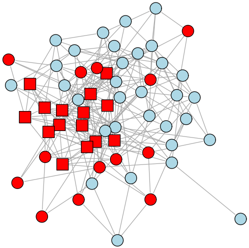
Zhao, Levina, and Zhu (2011b)

- Allow for background nodes that only have sparse links to other nodes
- Extract communities sequentially: at each step look for a set with a large number of links within and a small number of links to the rest of the network
- Stop when either the desired number is extracted or no more meaningful communities exist

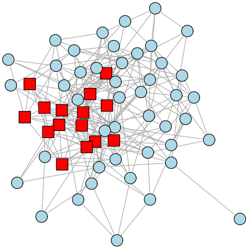
Toy example

- Block model with $K = 2$, $\pi_1 = 1/4$, $n = 60$, and
  $$P = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}.$$
- Compare partition into two communities (via modularity) to extraction of a single community
- Shapes represent the truth, colors represent estimation

Partition



Extraction



Extraction Criterion

Maximize

$$W(S) = \frac{O_{SS}}{n_S^2} - \frac{O_{SS'}}{n_S n_{S'}}$$

where $O_{SS} = \sum_{i,j \in S} A_{ij}$, $O_{SS'} = \sum_{i \in S, j \in S'} A_{ij}$.

- The links within the complement of set S do not matter.
- To avoid small communities, can use an adjusted criterion to encourage more balanced solutions:

$$W_a(S) = n_S n_{S'} \left(\frac{O_{SS}}{n_S^2} - \frac{O_{SS'}}{n_S n_{S'}} \right) .$$

Consistency of extraction

Theorem (Zhao, Levina, and Zhu 2011b)

Assume $K = 2$, WLOG $P_{11} \geq P_{22}$, and $P_{11} + P_{22} > 2P_{12}$. Both unadjusted and adjusted criteria are consistent under the block model.

Simulation I

- Two communities plus background, $n = 1000$
- Balanced ($n_1 = n_2 = 200$) and unbalanced ($n_1 = 100, n_2 = 200$)
- Generated from the block model with $K = 3$, $P_{12} = P_{23} = P_{13} = P_{33} = 0.05$
- Two levels of community strength: $P_{11} = 0.15$, $P_{22} = 0.12$, and $P_{11} = 0.20$, $P_{22} = 0.16$

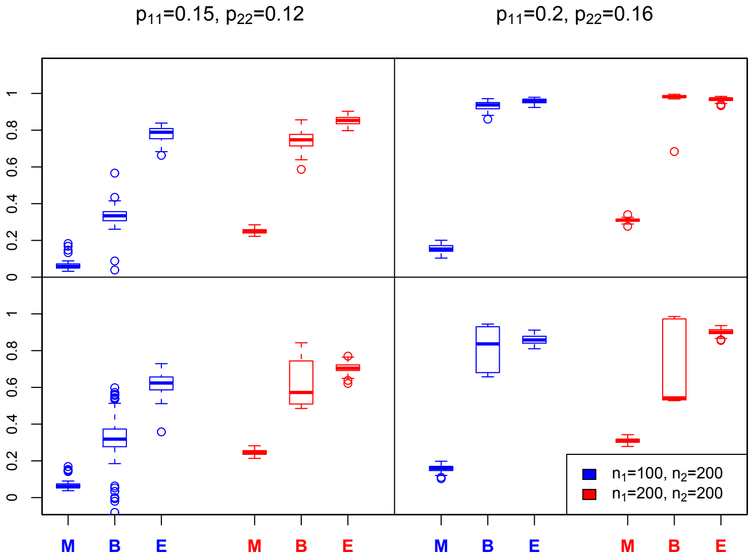
Simulation II

- Designed to test robustness to non-homogeneous degree distribution within communities

Simulation II

- Designed to test robustness to non-homogeneous degree distribution within communities
- Start with the same set-up as Simulation I
- In each community, double the degrees of the 10 highest-degree nodes by adding random edges to them in the same community
- Delete the same number of edges at random from all other edges in the same community

Results of simulations I (top) and II (bottom)



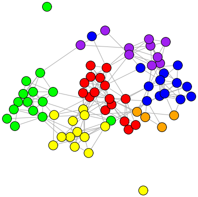
School friendship network

The school friendship network is compiled from the National Longitudinal Study of Adolescent Health (AddHealth) (http://www.cpc.unc.edu/projects/addhealth)

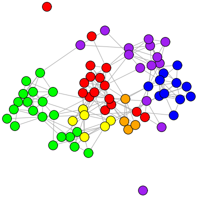
Grade 7: red
Grade 8: blue
Grade 9: green
Grade 10: yellow
Grade 11: purple
Grade 12: orange

Extraction on the school friendship network

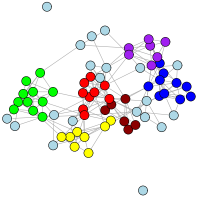
Grades



Modularity



Extraction



Future work

1. Determining the number of communities
2. Goodness-of-fit for network models

References

Y. Zhao, E. Levina, and J. Zhu. (2011a) Consistency of community detection in networks under degree-corrected stochastic block models. *Annals of Statistics.*, Volume 40, Number 4 (2012), 2266-2292.

Y. Zhao, E. Levina, and J. Zhu. (2011b) Community extraction for social networks. *Proc. Nat. Acad. Sci.*, 108(18):7321-7326.

Thank you!

Counter example

An example for the inconsistency of Erdos-Renyi modularity, block model likelihood and extraction.

$K = 2, \pi = (1/2, 1/2)$, and $P = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$.

$$\theta = \begin{cases} 1.6 & \text{w.p.} \frac{1}{2}, \\ 0.4 & \text{w.p.} \frac{1}{2}. \end{cases}$$

By grouping nodes with the same θ_i, the population values of ERM and BL are higher than the correct partition.

By extracting the nodes with high θ_i in a community, the population values of unadjusted and adjusted extract are higher than the correct extraction.

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:

(*) $F(H(R), \sum_{au} R_{.au})$ *is uniquely maximized over* $\{R : R \geq 0, \sum_k R_{kau} = \Pi_{au}\}$ *by* $R_{kau} = \Pi_{au}\delta_{ka}$ *for any* u*, where* $H \in \mathscr{R}^{K\times K}, R \in \mathscr{R}^{K\times K\times\infty}$*,* $H(R) = \sum_{abuv} x_u x_v P_{ab} R_{kau} R_{lbv}, R_{kau} = \frac{1}{n}\sum_{i=1}^{n} I(e_i = k, c_i = a, \theta_i = d_u)$.

Q is consistent under degree-corrected block models.

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