Consistency of Network Modularities

Peter Bickel

UC Berkeley

(Joint work with Aiyou Chen, *Bell Labs*)

Machine Learning Summer School, University of Chicago

June 11, 2009

(With assistance of Jing Lei)

Outline

1. The sub-community identification problem and modularities
2. A “nonparametric” probability model for unlabelled graphs and its relation to block models
3. The difficulties of maximum likelihood
4. Consistency of Newman-Girvan and other modularities and refined comparison
5. Algorithms, simulations and real data
6. Discussion

Community identification problem

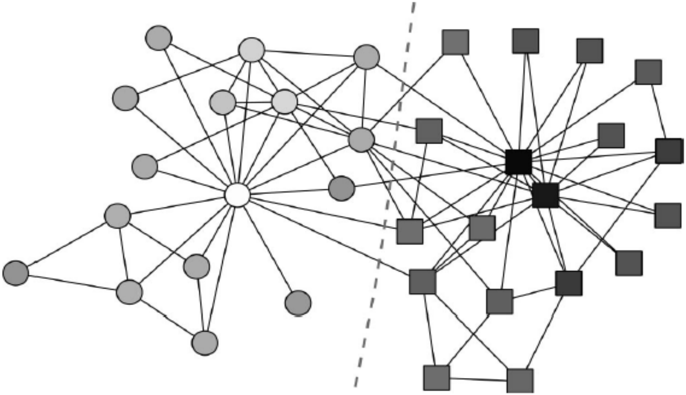

Fig. 1 Karate Club (Newman, PNAS 2007)

Ethnicity of students

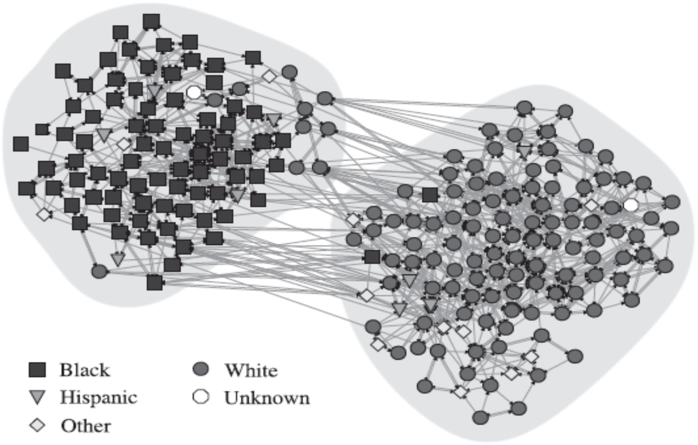

Fig. 2 Self-identified ethnicity of high school students (Newman & Leicht, PNAS 2007)

A Mathematical Formulation

- $G = (V, E)$: undirected graph
- $\{1, \cdots, n\}$: Arbitrarily labeled vertices
- A : adjacency matrix
- $A_{ij} = 1$ if edge between i and j (relationship)
- $A_{ij} = 0$ otherwise

The Problem

- $V = V_1 \cup \cdots \cup V_K$
- V_i : communities, $i = 1, \cdots, K$, where K is known.
- Problem: Determine V_j using only A

Approaches: Maximize Modularities

- Newman-Girvan modularity (Phys. Rev. E, 2004) $\mathbf{e} = (e_1, \cdots, e_n)$: $e_i \in \{1, \cdots, K\}$ (community labels)
- The modularity function:

$$Q_N(\mathbf{e}) = \sum_{k=1}^{K} \left(\frac{O_{kk}(\mathbf{e}, A)}{D_+} - \left(\frac{D_k(\mathbf{e})}{D_+} \right)^2 \right),$$

where

$$\begin{aligned} O_{ab}(\mathbf{e}, A) &= \sum_{i,j} A_{ij} \mathbf{1}(e_i = a, e_j = b) \\ &= (\# \text{ of edges between } a \text{ and } b) \quad a \neq b \\ &= 2 \times (\# \text{ of edges between members of } a), \quad a = b \\ D_k(\mathbf{e}) &= \sum_{l=1}^{K} O_{kl}(\mathbf{e}, A) \\ &= \text{sum of degrees of nodes in } k \\ D_+ &= \sum_{k=1}^{K} D_k(\mathbf{e}) = 2 \times (\# \text{ of edges between all nodes }) \end{aligned}$$

Block Models (Holland, Laskey and Leinhardt 1983)

Probability models

- Community label: $\mathbf{c} = (c_1, \cdots, c_n)$ i.i.d. multinomial $(\pi_1, \cdots, \pi_K)$.
- Relation:

$$\mathbb{P}(A_{ij} = 1 | c_i = a, c_j = b) \quad = \quad P_{ab}.$$

- A_{ij} conditionally independent

$$\mathbb{P}(A_{ij} = 0) \quad = \quad 1 - \sum_{1 \le a, b \le K} \pi_a \pi_b P_{ab}.$$

Frequentist Approach to Sub-community Identification

(Similar for Bayesian)

The problem is NP complete but in practice seems solvable quickly.

Solution

1) Compute likelihood (A, π, P)

2) Maximize by EM

3) Compute $\hat{P}, \hat{\pi}$, $\mathbb{P}(c_i = a | A, \hat{P}, \hat{\pi})$

4) Classify by maximizing

$$P(c_i = a | A, \hat{P}, \hat{\pi}).$$

The Swapping Algorithm

Iterate the loop below until modularity stops increasing:

For $i = 1 : n$

- Label switching: $e_i = \arg\max Q(\{e_1, \ldots, e_n\})$.
- Compute modularity: $Q(\{e_1, \ldots, e_n\})$.

For profile likelihood, one can mix label switching with parameter update to speed up computation.

Nonparametric Asymptotic Model for Unlabeled Graphs

Aldous/Kallenberg (1983, 2005)

$$\mathcal{L}(A_{ij} : i,j \geq 1\} \quad = \quad \mathcal{L}(A_{\pi_i,\pi_j} : i,j \geq 1),$$

for all permutations $\pi \iff$

$$\exists \quad g : [0,1]^4 \to \{0,1\} \text{ such that } A_{ij} \quad = \quad g(\alpha, \xi_i, \xi_j, \eta_{ij}),$$

where

α, ξ_i, η_{ij}, all i, $j \geq i$, i.i.d. $\mathcal{U}(0,1)$, $g(\alpha, u, v, w) = g(\alpha, v, u, w)$,

$\eta_{ij} = \eta_{ji}$.

Ergodic Models

$\mathcal{L}$ is an ergodic probability iff for g with $g(u, v, w) = g(v, u, w)$ $\forall(u, v, w)$,

$$A_{ij} = g(\xi_i, \xi_j, \eta_{ij}).$$

$\mathcal{L}$ is determined by

$$h(u, v) \equiv \mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v)$$

$$h(u, v) = h(v, u).$$

Identifiability

- h is not uniquely determined.

  If $\psi : [0,1] \to [0,1]$ has $\psi(\xi) \sim \mathcal{U}(0,1)$, then

  $\{h(\psi(\xi_i), \psi(\xi_j)), i,j \geq 1\}$ has the same law $\mathcal{L}$.

- Proposal

$$g(u) \equiv \mathbb{P}(A_{ij} = 1 | \xi_i = u) = \int_0^1 h(u,v)dv.$$

$h_{\mathrm{CAN}}(u,v)$ is uniquely determined by:

i) $\{h_{\mathrm{CAN}}(\xi_i, \xi_j)\} \sim \mathcal{L}$.

ii) $g_{\mathrm{CAN}}(u) = \int_0^1 h_{\mathrm{CAN}}(u,v)dv \uparrow$ in u.

Identifiability (cont'd)

- $g(\xi_i) \sim F$, which is identifiable. Assume that F is continuous.
- Define measure-preserving

$$\psi(u) = F(g(u)).$$

- $h_{\text{CAN}}(u, v) = h(\psi(u), \psi(v))$

  $g_{\text{CAN}}(u) = F^{-1}(u)$, $\uparrow$.

- If $h(\xi_i, \xi_j) \sim \mathcal{L} \sim h_{\text{CAN}}(\xi_i, \xi_j)$, then

$$P[g(\xi) \leq u] = F(u) = g^{-1}(u),$$

if $g \uparrow$.

Block Models as Approximations

- $h(u, v) = h_{ab}$
  if $\sum_{i=1}^{a-1} \pi_i \le u < \sum_{i=1}^{a} \pi_i$ and $\sum_{i=1}^{b-1} \pi_i \le v < \sum_{i=1}^{b} \pi_i$,
  $\sum_b \pi_b h(a, b) \uparrow$ in a.
- Sieve: Fit K_n blockmodel $\lambda_1 = \cdots = \lambda_K = K^{-1}$.

$$\hat{h}_n(u, v) \;=\; \hat{h}_{ab}$$

for $\frac{a-1}{K} \le u < \frac{a}{K}$ and $\frac{b-1}{K} \le v < \frac{b}{K}$.

Profile Likelihood Modularities

$$Q_L(\mathbf{e}) = \sum_{a=1}^{K} n_a(\mathbf{e}) \log \frac{n_a(\mathbf{e})}{n} + \sum_{a<b} n_{ab}(\mathbf{e}) \tau \left(\frac{O_{ab}(\mathbf{e}, A)}{n_{ab}(\mathbf{e})} \right) + \sum_{a=1}^{K} n_{aa}(\mathbf{e}) \tau \left(\frac{O_{aa}(\mathbf{e}, A)}{2n_{aa}(\mathbf{e})} \right)$$

where

$$\tau(x) = x \log x + (1-x)\log(1-x).$$

$$\mathbf{e} \equiv (e_1, \ldots, e_n).$$

$$n_a(\mathbf{e}) \equiv \sum \mathbf{1}(e_i = a).$$

$$n_{ab}(\mathbf{e}) = \begin{cases} n_a(\mathbf{e}) n_b(\mathbf{e}), & a \neq b \\ \binom{n_a}{2}, & a = b \end{cases}.$$

Asymptotic Approximation for Block Models

Consider sequence of models $\mathcal{L}_n$ (single model if $\lambda_n = n$).

$$\mathbb{E}D_+ = \gamma_n \equiv \gamma n \lambda_n$$

$$\gamma_n = n^2 \sum_{a,b} \pi_a \pi_b P_{ab}^{(n)}$$

$$\mathcal{L}_n \equiv \{\mathcal{L}(A : \pi, P^{(n)}) : P^{(n)} = n^{-1} W \lambda_n\}, \quad W = ||W||_{ab}, W_{ab} > 0.$$

$$\lambda_n \to \infty$$

$$\lambda_n \leq n$$

Consistency of Modularities

General Modularity:

- Given Q: $K \times K$ positive matrices $\times K$ simplex $\rightarrow \mathbb{R}^+$.
- Q modularity $\equiv Q\left(\frac{O(\mathbf{e},A)}{D_+}, \mathbf{f}(\mathbf{e})\right)$.

  $O(\mathbf{e}, A) \equiv ||o_{ab}(\mathbf{e})||$, $\mathbf{f}(\mathbf{e}) \equiv (f_1(\mathbf{e}), \ldots, f_K(\mathbf{e}))^T$, $f_j(\mathbf{e}) \equiv \frac{n_j}{n}$.

  $\hat{\mathbf{c}} \equiv \arg\max Q\left(\frac{O(\mathbf{e},A)}{D_+}, \mathbf{f}(\mathbf{e})\right)$.

Conditions

C1: a) The matrix W has no two rows equal and all elements are > 0.

b) $\pi_i > 0$, $i = 1, \dots, K$.

C2: $\mathcal{M} = \{R : R_{ab} \geq 0, \text{ all } a, b, R^T \mathbf{1} = \pi\}$.

$F(R) \equiv Q(RP^{(n)}R^T, R\mathbf{1})$, $P^{(n)} = \frac{\lambda_n}{n} W$.

$F(R)$ is uniquely maximized over $\mathcal{M}$ at

$R = \pi^D \equiv \mathrm{diag}(\pi_1, \dots, \pi_K)$.

C3: $\frac{\partial F}{\partial r_{ab}} \left(\pi^D\right) < 0$, $a \neq b$.

C4: $\frac{\lambda_n}{\log n} \to \infty$.

Global Consistency

Suppose $\hat{\mathbf{c}} = \arg\max Q\left(\frac{O(\mathbf{e},A)}{D_+}, \mathbf{f}(\mathbf{e})\right)$

and C1-4 hold. Then:

$$\frac{1}{\lambda_n} \log P[\hat{\mathbf{c}} \neq \mathbf{c}] \to -s_Q, \quad \text{with } s_Q > 0.$$

See also Snijders and Nowicki (1997) Journal of Classification.

Estimation of Block Model Parameters

- Identify $P^{(n)}$ and P. Assume C1-C4.

$$\begin{aligned}
\hat{\pi}_a &\equiv \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{1}(\hat{c}_i = a) \equiv \tfrac{\hat{n}_a}{n} \\
\hat{P}_{ab} &= \tfrac{1}{\hat{n}_a \hat{n}_b} \textstyle\sum_{i,j}\{\mathbf{1}(\hat{c}_i = a, \hat{c}_j = b)\} \quad a \neq b \\
\hat{P}_{aa} &= \tfrac{1}{\binom{\hat{n}_a}{2}} \textstyle\sum_{i<j}\{\mathbf{1}(\hat{c}_i = a, \hat{c}_j = a)\}
\end{aligned}$$

- Efficiency: If $\lambda_n \sim n$,

$$\begin{aligned}
\sqrt{n}(\hat{\pi}_a - \pi_a) &\Rightarrow N(0, \pi_a(1-\pi_a)) \\
n(\hat{P}_{ab} - P_{ab}) &\Rightarrow N\left(0, \tfrac{P_{ab}(1-P_{ab})}{\pi_a \pi_b}\right), \quad a \neq b \\
&\Rightarrow N\left(0, \tfrac{2P_{aa}(1-P_{aa})}{\pi_{aa}^2}\right), \quad a = b
\end{aligned}$$

- If $\lambda_n/n \to 0$ then this is really a statement about estimating W at rate λ_n^{-1}.

Newman Girvan (NG)

- Let $\Delta = R - \pi^D$

  $\mathcal{E} = P - (\pi^T P \pi)^{-1} P \pi \pi^T P$

  $G_N(\Delta) = \mathrm{trace}(\Delta \mathcal{E} \Delta^T + \Delta \mathcal{E} \pi^D + \pi^D \mathcal{E} \Delta^T)$.

- C2 for NG $\equiv G_N(\Delta)$ uniquely maximized by $\Delta = 0$ on

  $\{\Delta : \Delta^T \mathbf{1} = 0, \Delta_{ab} \geq 0, a \neq b\}$.

Result

1) NG satisfies C2, C3 if $\mathcal{E}$ has all diagonal entries positive and all nondiagonal entries negative.

2) But C2, C3 and consistency may fail.

Newman-Girvan NG: counter examples

$K = 3, \pi = (.6, .1, .3)$ and

$$P = \begin{bmatrix} 0.8 & 0.1 & 0.3 \\ 0.1 & 0.7 & 0.6 \\ 0.3 & 0.6 & 0.9 \end{bmatrix}$$

With true labeling, NG modularity approaches 0.10. However, the true maximum modularity, equal 0.12, is achieved by merging the last two communities.

If $U_1, \ldots, U_m$ are independent, $EU_j = 0, \quad |U_j| \leq 1, \quad 1 \leq j \leq m$, $S_m = \sum_{j=1}^m U_j$, then

$$\frac{1}{m} \log P\left[\left|\frac{S_m}{m}\right| \geq c\right] \sim \rho < 0,$$
$$P\left[\left|\frac{S_m}{m}\right| \geq t\right] \leq Ce^{-m\frac{t^2}{2}}, \quad \frac{t}{\sqrt{m}} \to \infty,$$
$$\max_{1 \leq l \leq L}\left|\frac{S_{ml}}{m}\right| = O_P\left(\sqrt{\frac{\log L}{m}}\right) = o_P(1)$$

if $L \sim K^{\sqrt{m}}$

In our case $m \sim n^2 \sim \#$ of edges.

$$L = 2^n.$$

Profile Likelihood

$$F(R) = \sum_{k<l} (R\mathbf{1}\mathbf{1}^T R^T)_{kl} \tau \left(\frac{(RPR^T)_{kl}}{(R\mathbf{1}\mathbf{1}^T R^T)_{kl}} \right) + \sum_{1 \le k \le K} \frac{1}{2} (R\mathbf{1}\mathbf{1}^T R^T)_{kk} \tau \left(\frac{(RPR^T)_{kk}}{(R\mathbf{1}\mathbf{1}^T R^T)_{kk}} \right)$$

$\tau(x) \equiv x \log x + (1-x) \log(1-x)$

PL. is consistent under C1.

Simulation

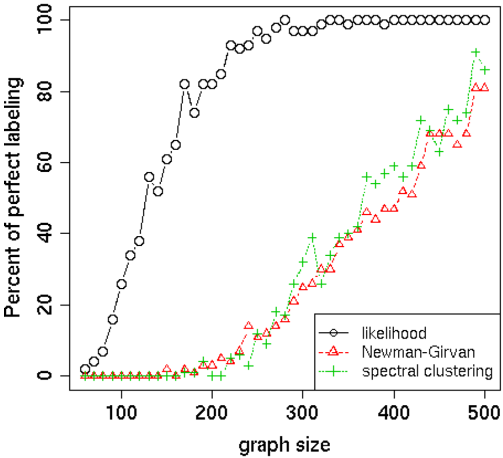

Fig. 3 Compare NG, spectral clustering and profile likelihood with $K = 3$.

Real Data: Private Branch Exchange

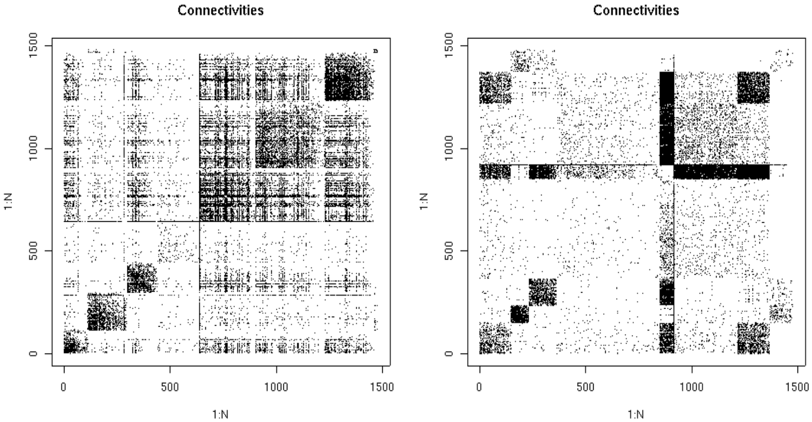

Fig.4 Different communities formed by NG and profile likelihood

Quicker Optimal Approach

1. Estimate

$$\hat{P}^{(n)} = P^{(n)} + O_P(n^{-1})$$
$$\hat{\pi} = \pi + O_P(n^{-1/2})$$

   by MOM, using degree, triangles, etc.

2. Let

$$\begin{aligned} \hat{f}_i(k, A) &\propto P[c_i = k | A, \hat{P}, \hat{\lambda}], \\ \hat{c}_i &= \arg\max \hat{f}_i(k, A). \end{aligned}$$

3. Computing $\hat{f}_i$

   Simulate $\mathbf{c}_s = (c_{s1}, \ldots, c_{sn})^T$, $s = 1, \ldots, S$. c_{si} iid $\mathcal{M}(1, \hat{\pi})$,

   $i = 1, \ldots, n$.

$$\hat{f}_{is}(k,A) = \prod_{a \le b} \hat{P}_{ab}^{O_{abs}^{-i,k}} (1 - \hat{P}_{ab})^{n_{abs}^{-i,k} - O_{abs}^{-i,k}}$$

$$\hat{f}_i(k,A) = \frac{1}{S} \sum_{s=1}^{S} \hat{\pi}_k \hat{f}_{is}(k,A)$$

$$
\begin{aligned}
O_{abs}^{-i,k} &= \textstyle\sum_{r,t} A_{rt} \left(1 - \frac{\delta_{ab}}{2}\right) \mathbf{1}(c_{sr}^{-i,k} = a, c_{st}^{-i,k} = b) \\
n_{abs}^{-i,k} &= n_{as}^{-i,k} n_{bs}^{-i,k}, \quad a \neq b \\
&= \tbinom{n_{as}^{-i,k}}{2}, \quad a = b \\
n_{as}^{-i,k} &= \textstyle\sum_{j=1}^{n} \mathbf{1}(c_{sj}^{-i,k} = a) \\
c_{sj}^{-i,k} &= c_{sj}, \quad j \neq i \\
&= k, \quad j = i
\end{aligned}
$$

Discussion

1. The problem of $\lambda_n = n$ is easy. Given $\hat{\mathbf{c}}$, estimation of P, π is easy. Just count.
2. Framework lends itself to generalization as in ordinary statistics. Introduce covariates $\{Z_i\}$ or $\{Z_{ij}\}$. Eg, location, # of telephone calls per week between i and j.
3. Generalization to directed case based on characterization.
$$\begin{aligned} X_{ij} &\neq X_{ji} \leftrightarrow g(w,u,v,z) \text{ not necessarily symmetric in } (u,v) \\ \tilde{X}_{ij} &\equiv X_{ij} + X_{ji} \end{aligned}$$
Estimate directions after communities. Fit undirected graph.
4. Nonparametric theory.

   Estimate affinity function $h_{\mathrm{CAN}}(u,v)$.