# *Topics in Network Models*

MSU, Sepember, 2012


Peter Bickel

*Statistics Dept. UC Berkeley*

(Joint work with S. Bhattacharyya *UC Berkeley*, A. Chen *Google*, D.

Choi *UC Berkeley*, E. Levina *U. Mich* and P. Sarkar *UC Berkeley*)

# *Outline*

1. Networks: Examples, References
2. Some statistical questions
3. A nonparametric model for infinite networks.
4. Asymptotic theory for $\frac{\lambda_n}{\log n} \to \infty$, $\lambda_n =$ Average degree.
5. Results of B. and Chen (2009) on block models.
6. Consequences for maximum likelihood and variational likelihood.
7. Count statistics, bootstrap and $\lambda_n = O(1)$.

# *I.Identifying Networks and II.Working With Given Ones*

*I* Given vectors of measurements $\mathbf{X_i}$, for example, given gene expression sequence nearby binding site information, physical (epigenetic information), protein assays etc. Determine dependency/causal relation between genes.

Tools: Gaussian graphical models, clustering etc.

*II* Given a network of relations (edges) identify higher level structures, clusters, like pathways in genomics. In practice, both can be done simultaneously. Focus on the models for II, in the simplest case of unlabelled graphs.

# *Example: Social Network*

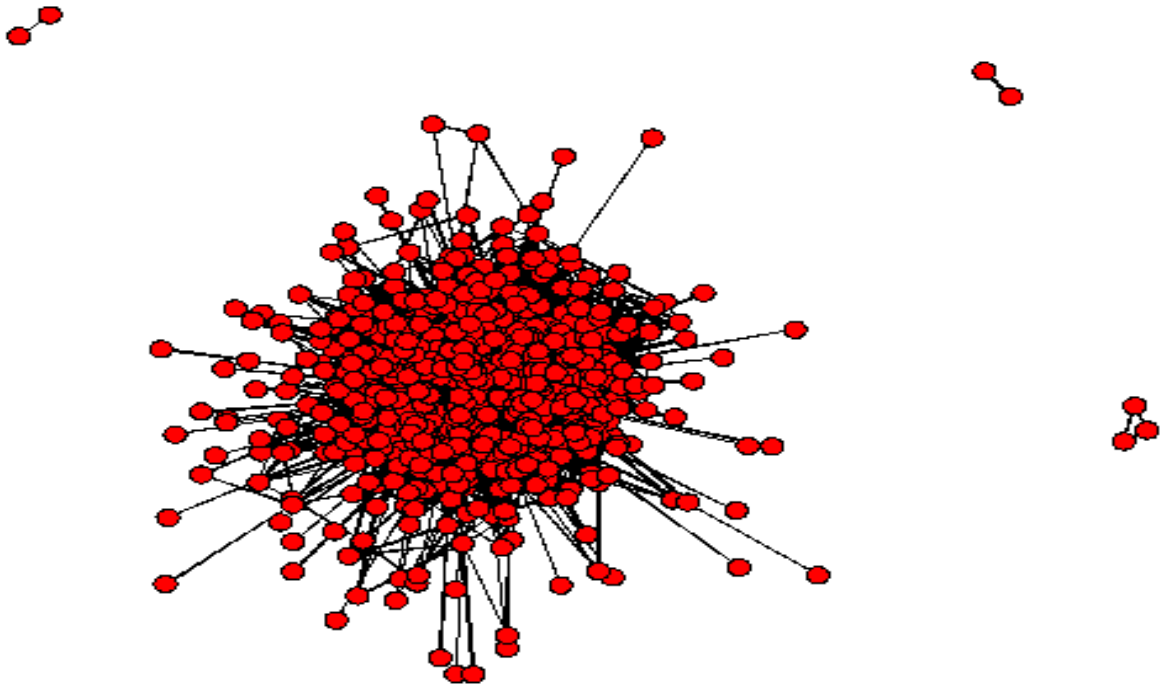

*Figure:* Facebook Network for Caltech with 769 nodes and average degree 43.

# *Example: Bio Network*

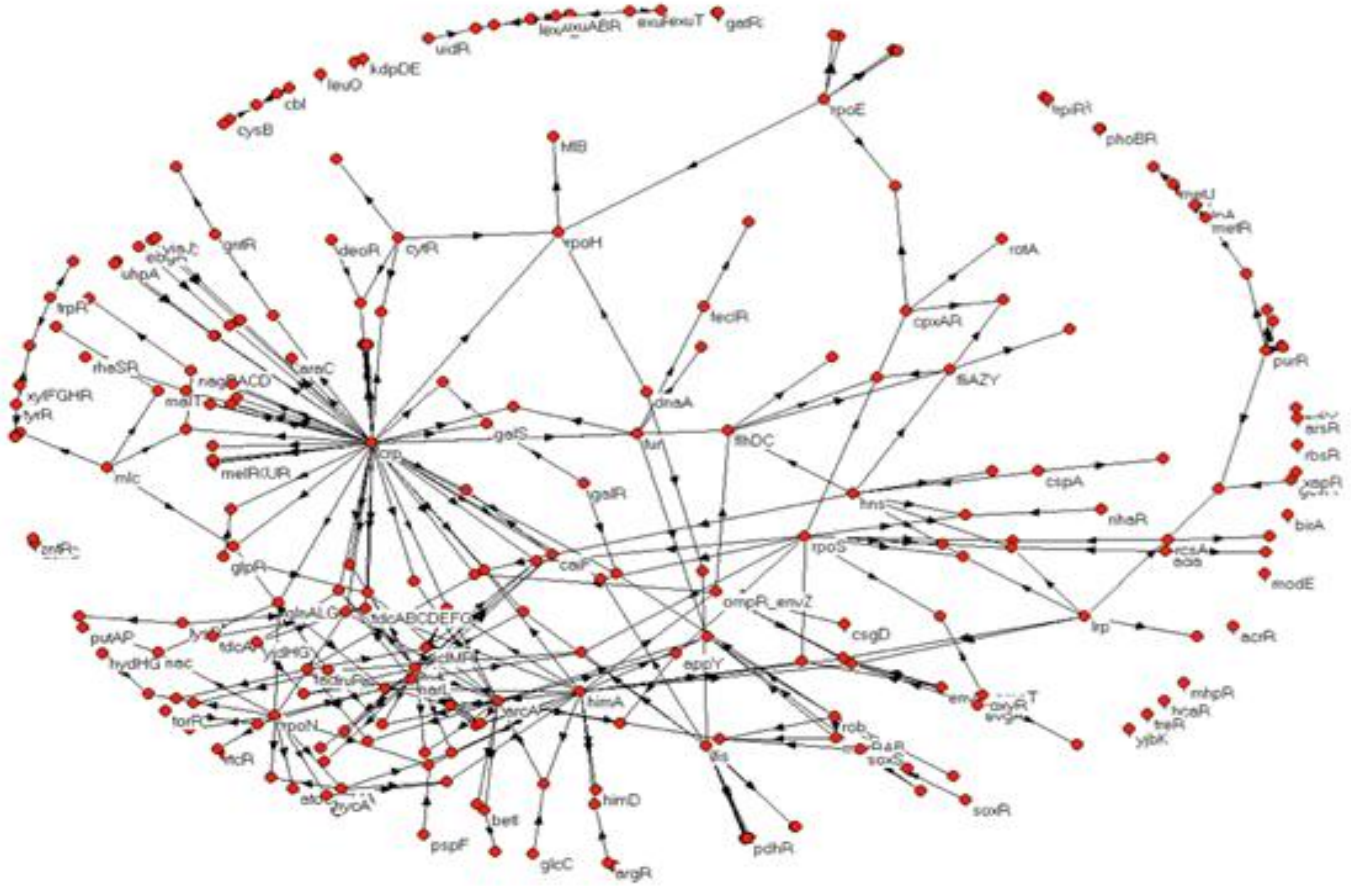


*Figure:* Transcription network of E. Coli with 423 nodes and average degree 2.45.

# *Example: Collaboration Network*

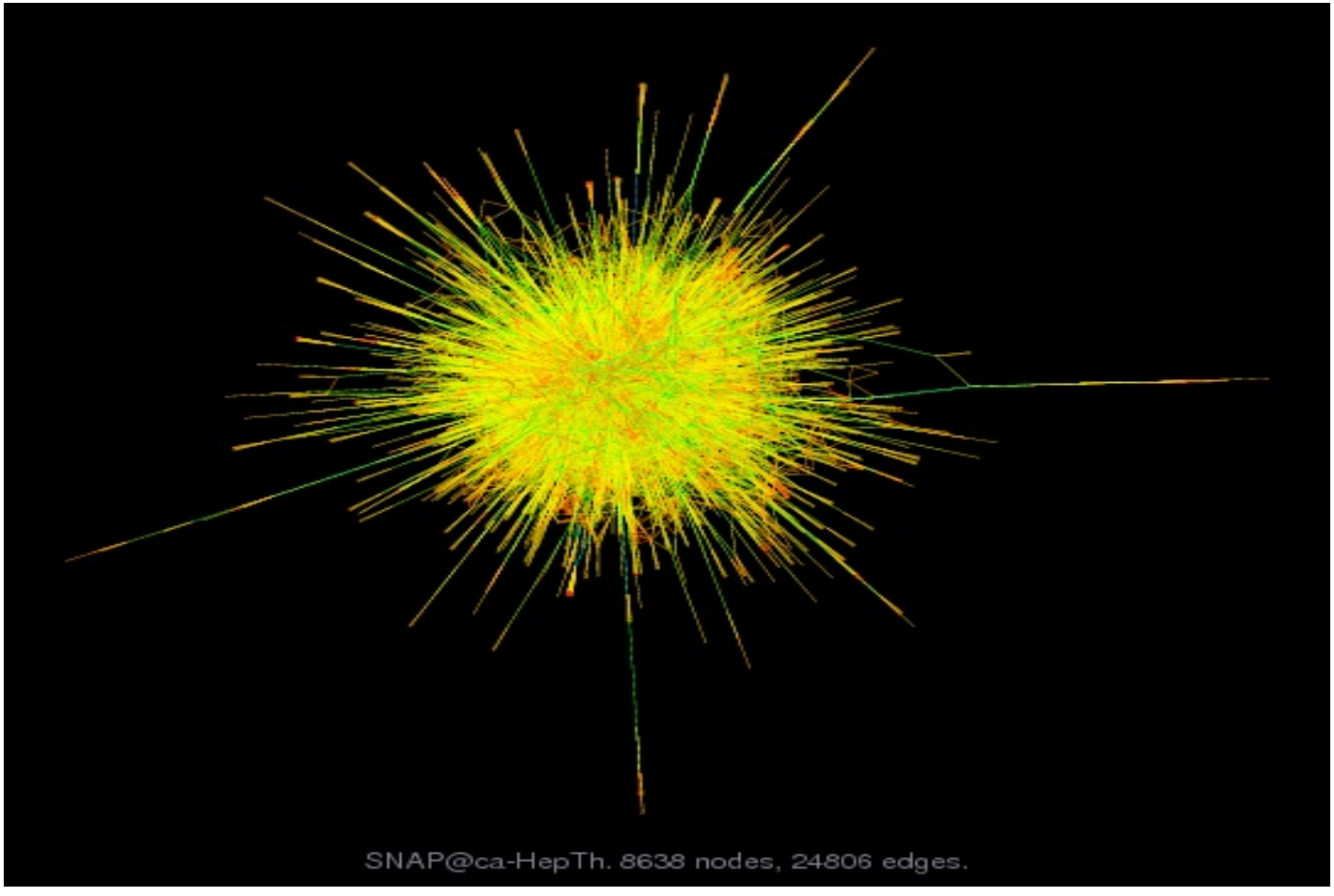


*Figure:* Collaboration Network in Arxiv for High Energy Physics with 8638 nodes and average degree 5.743.

# *References on Networks*


- **Books**
  1. (Phys/CS) M.E.J. Newman (2010) Networks: An introduction. Oxford
  2. (Stat) Eric D. Kolaczyk (2009) Statistical Analysis of Network Data
  3. (Soc. Science/CS) David Easley and Jon Kleinberg (2010) Networks, crowds and markets: Reasoning about a highly connected world. Cambridge University Press
  4. (Soc. Science) K. Faust and S. Wasserman (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, New York.
  5. (Prob/CS) Fan Chung, Linyuan Lu (2004) Complex graphs and networks. CBMS # 107 AMS
- **Papers**
  1. (Prob/Math) Bela Bollobas, Svante Janson, Oliver Riordan (2007) The Phase Transition in Random Graphs. Random Structures and Algorithms, 31 (1) 3-122
  2. (Prob/Math) Oliviera RI (2010): Concentration of Adjacency Matrix and Graph Laplacian for Graphs with Independent Edges. Arxiv 0911.0600.
  3. (Stat) Celisse, Daudin and Pierre (2011): Consistency of ML in Block Models. Arxiv
  4. (Stat) Daudin, Picard and Robin (2008): A Mixture Model for Random Graphs. J. Stat. Comp.
  5. (Stat) Chaudhuri, K., Chung F., Tsiatis A. (2012). Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model. JMLR.
  6. (Stat) Liben-Nowell, D., and Kleinberg, J., (2007). Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology.
- **Our works**
  1. B. and A. Chen (2009) A nonparametric view of network models and Newman-Girvan and other modularities, PNAS
  2. B., E. Levina and A. Chen (2011) Method of Moments Estimation For Networks, Ann. Stat.
  3. B. and D. Choi (2012) , Arxiv.

# *Questions*

**Focus**

*(i)* Community identification for block models (simple and degree-corrected).

**Other**

*(ii)* Link Prediction: Predicting edges between nodes based on partially observed graph. (Liben-Nowell and Kleinberg (2007)), Zhao and Levina (2012).

*(iii)* Model selection: Number of blocks.

*(iv)* Testing stochastic identity of graphs (networks) using count statistics (motifs) (Wasserman and Faust (1994)).

*(v)* Error bars on descriptive statistics, eg. homophily.

*(vi)* Model incorporating edge identification errors.

*(vii)* Directed graphs: graphs with additional edge and vertex information.

# *Erdős-Rényi and Block Models (Holland, Laskey and Leinhardt 1983)*

- Graph on $n$ vertices is **equivalent** to $A_{n\times n}$ adjacency matrix.
- Probability on $\mathcal{S}_n = \{\text{all } n \times n \text{ symmetric 0-1 matrices}\}$. (For undirected graphs)
- **E-R:** $A_{ij} \overset{i.i.d.}{\sim} Ber(p)$.
- **H-L-L:** $K$ "communities"
  - $c_a \equiv \mathbf{1}[\text{vertex } i \in a]$
  - $\mathbf{c} \sim \mathcal{M}(n, (\pi_1, \ldots, \pi_K))$
  - Given $\mathbf{c}$, $A_{ij}$ are independent and

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

where, $\sum_{a,b} w_{ab} = 1$ and $w_{ab} = w_{ba}$.

# *Nonparametric Asymptotic Model for Unlabeled Graphs*

Given: $P$ on $\infty$ graphs

Aldous/Hoover (1983)

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

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

$\alpha, \xi_i, \eta_{ij}$, 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

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

Notes:

- Equivalent: Hoff, Raftery, Handcock (2002). JASA
- More general: Bollobas, Riordan, Janson (2007). Random Structures and Algorithms

# *"Parametrization" of NP Model*

- $h$ is not uniquely defined.
- $h\Big(\varphi(u), \varphi(v)\Big)$, where $\varphi$ is measure-preserving, gives same model.
- **Property CAN** holds If there exists $h_{CAN}$ such that

$$\tau(z) \equiv \int_0^1 h_{CAN}(z,u)du = \mathbb{P}[A_{ij} = 1 | \xi_i = z]$$

is

(a) monotonically non-decreasing and

(b) If $F(\cdot)$ is CDF of $\tau(\xi)$, $w\left(F\tau(\xi_1), F\tau(\xi_2)\right) \sim w(\xi_1, \xi_2)$.

- If $\tau$ is strictly increasing, such an $h_{CAN}$ always exists.
- $\xi_i$ could be replaced by any continuous variables or vectors - but there is no natural unique representation.

# *Asymptotic Approximation*

- $h_n(u,v) = \rho_n w_n(u,v)$
- $\rho_n = \mathbb{P}[\text{Edge}]$
- $w(u,v)dudv = \mathbb{P}\left[\xi_1 \in [u, u+du], \xi_2 \in [v, v+dv] | \text{Edge}\right]$
- $w_n(u,v) = \min\left\{w(u,v), \rho_n^{-1}\right\}$
- Average Degree $= \frac{E(D_+)}{n} \equiv \lambda_n \equiv \rho_n(n-1)$.
- $\lambda_n$ ranges from $O(1)$ to $O(n)$.

# *Examples of models*

*(I)* **Block models:**

$$h_{CAN}(u,v) \equiv F_{ab} \equiv \frac{w_{ab}}{\pi_a \pi_b} \text{ on } I_a \times I_b,\ |I_a| = \pi_a,\ a = 1, \ldots, K.$$

**Degree-Corrected:** (Karrer and Newman (2010))

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

$\theta_i$ is independent of $\mathbf{c}$, $\mathbb{P}[\theta_i = \lambda_j] = \rho_j$, $\sum_{j=1}^{J} \rho_j 1$.

Parametric sub model of $M = KJ$ block model.

**Not to be considered**

*(II)* **Preferential Attachment:** (De Solla-Price (1965))

(Asymptotic version, $w$ not bounded)

$$w(u,v) = \frac{\tau(u)}{\int_u^1 \tau(s)ds} \mathbf{1}(u \leq v) + \frac{\tau(v)}{\int_v^1 \tau(s)ds} \mathbf{1}(v \leq u)$$

$$\tau(u) = \int_0^1 h(u,v)dv$$

Dynamically defined, $\tau(u) \sim (1-u)^{-1/2}$ is equivalent to power law of *degree distribution* $F \equiv \tau^{-1}$.

# *Two Questions*

*(I)* *(a)* Estimate $\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{F}$.

*(b)* Classify vertex $i$ by its community.

*(II)* Given $i, j$ construct $\hat{w}(\xi_i, \xi_j)$ ($\hat{w}$ is an estimate of $w$) to predict whether $A_{ij} = 1$.

# *3 Regimes*

*(a)* If $\frac{\lambda_n}{\log n} \to \infty$, equivalent to,

$\mathbb{P}$[there exists an isolated point] $\to 0$.

*(b)* If $\lambda_n \to \infty$, full identifiability of $w$.

*(c)* If $\lambda_n = O(1)$, phase boundaries and partial identifiability of $w$.

# *Block Models: Community Identification and 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

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

# *Profile Likelihood*

- Given $\mathbf{e}$ estimate parameters and plug into block model likelihood.
- Always consistent for classification and efficient estimation.

# *Global Consistency*

**Theorem 1**

If population version of $F$ is "consistent" and $\frac{\lambda_n}{\log n} \to \infty$, then

$$\limsup_n \lambda_n^{-1} \log \mathbb{P}\left[\hat{\mathbf{c}} \neq \mathbf{c}\right] \leq -s_Q, \text{ with } s_Q > 0.$$

Extension to $F_n \approx F$ requires simple condition. See also Snijders and Nowicki (1997) J. of Classification.

# *Basic Approach to Proof*

General Modularity: Data and population

- Given $Q_n$: $K \times K$ positive matrices $\times K$ simplex $\longrightarrow \mathbb{R}^+$.
- $Q_n(\mathbf{e}, A) = F_n\left(\frac{O(\mathbf{e},A)}{\mu_n}, \frac{D_+}{\mu_n}, f(\mathbf{e})\right)$.

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

  $\hat{\mathbf{c}} \equiv \arg\max Q_n(\mathbf{e}, A)$.

  $\mu_n = E(D_+) = (n-1)\lambda_n$.
- NG: $F_n \equiv F$.
- Population: Replace random vectors by expectations under block model

## Corollary

Under the given conditions if

$$\hat{\pi}_a = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{c}_i = a) \equiv \frac{\hat{n}_a}{n},$$
$$\hat{W} = \frac{O(\hat{c},A)}{D_+},$$

then if $S = \Delta^{-1} W \Delta^{-1}$, where, $\Delta = \text{Diag}(\boldsymbol{\pi})$,

$$\sqrt{n}(\hat{\pi} - \pi) \Rightarrow \mathcal{N}(\mathbf{0}, \pi^D - \pi\pi^T),$$
$$\sqrt{n\lambda_n}(\hat{S} - S) \Rightarrow \mathcal{N}\left(\mathbf{0}, \Sigma(\pi, W)\right).$$

These are efficient.

# *Maximum Likelihood, Variational Likelihood for Block Models*

(B. and Choi (2012) Arxiv, Celisse et. al. (2011) Arxiv)

- **Complete data likelihood for block data**

$$f(\mathbf{z}, A, \theta) = \prod_{i=1}^{n} \pi_{z_i} \prod_{i \leq j} \left(\rho S_{z_i z_j}\right)^{A_{ij}} \left(1 - \rho S_{z_i z_j}\right)^{1 - A_{ij}}$$

$\mathbb{P}[z_i = a] = \pi_a$ and $z_1, \ldots, z_n$ i.i.d.

- **Graph likelihood ratio**

$$\frac{g}{g_0}(A, \theta) = \mathbb{E}_0 \left( \frac{f}{f_0}(\mathbf{z}, A, \theta) \mid A \right)$$

where, $f_0 \equiv f(\mathbf{z}, A, \theta_0)$, $g_0 \equiv g(A, \theta_0)$ and $\theta \equiv (\rho, \boldsymbol{\pi}, S)^T$.

# *Variational Likelihood (Daudin et. al. (2009))*

Define,

$$q(\mathbf{z}, \tau) \equiv \prod_{i=1}^{n} \tau_i(z_i)$$

$$f_{VAR}(\mathbf{z}, A, \theta) = q(\mathbf{z}, \tau(A, \theta), \theta) g(A, \theta)$$

$$\tau(A, \theta) = \arg\min_{\tau} D\left(q(\mathbf{z}, \tau) \mid f(\mathbf{z}|A, \theta)\right)$$

where, $D(f_1|f_2) = \int \left(\log \frac{f_1}{f_2}\right) f_1 d\mu$.

Notation:

- $\hat{\hat{\theta}}$: Complete Data likelihood estimator,
- $\hat{\theta}$: Graph Likelihood estimator,
- $\hat{\theta}_{VAR}$: Variational Likelihood estimator.

# *Theorem*

**Theorem (B. and Choi)**

If $\frac{\lambda_n}{\log n} \to \infty$, there exists $E$ such that $\sup_\theta \mathbb{P}_\theta(\bar{E}|A) \xrightarrow{P_0} 0$ such that

*(a)*

$$\frac{g}{g_0}(A, \theta)\mathbf{1}_E(\mathbf{z}, A) = \frac{f}{f_0}(\mathbf{z}, A)(1 + o_{\mathbb{P}}(1))\mathbf{1}_E(\mathbf{z}, A)$$

*(b)* The same holds for $\frac{f_{VAR}}{f_{0,VAR}}$

*(c)* $f_{VAR}$ can be used for consistent classification.

# *Consequence of Theorem*

Hence,

- $\sqrt{n}(\hat{\hat{\pi}} - \hat{\pi}) = o_{\mathbb{P}}(1)$.
- $\sqrt{n}(\hat{\hat{\pi}} - \hat{\pi}_{VAR}) = o_{\mathbb{P}}(1)$.
- $\sqrt{n\lambda_n}(\hat{\hat{S}} - \hat{S}) = o_{\mathbb{P}}(1)$.
- $\sqrt{n\lambda_n}(\hat{\hat{S}} - \hat{S}_{VAR}) = o_{\mathbb{P}}(1)$.
- $\sqrt{n\lambda_n}(\frac{\hat{\rho}}{\rho} - 1) = O_{\mathbb{P}}(1)$.
- $\hat{\rho} \equiv \frac{1}{n}\sum_{i=1}^{n} D_i$, $D_i = \sum_j A_{ij}$.

# *Concentration Results and Consequences in Regimes (a) and (b)*

**Theorem (B., Levina and Chen (2012)), (Channarond, Daudin, Robin (2012))**

*(i)* If $\lambda_n \to \infty$

$$\left| \frac{D_i}{\bar{D}} - \tau(\xi_i) \right| = O_P(\lambda^{-1/2})$$

*(ii)* If $\frac{\lambda_n}{\log n} \to \infty$

$$\max_i \left| \frac{D_i}{\bar{D}} - \tau(\xi_i) \right| = o_P(1)$$

# *Statistical Consequences of (ii)*

(Channarond, Daudin, Robin (2011, Arxiv))

- Let us suppose that the block model has the Property CAN with parameters, $(\boldsymbol{\pi}, W, \rho)$, where, $W_{ab} = \mathbb{P}[i \in a, j \in b | ij \text{ is an Edge}]$.
- **Algorithm**: Apply $k$-means to $\left\{\frac{D_i}{\bar{D}}\right\}$ $1 \leq i \leq n$.
- If $\lambda_n \to \infty$ and $\hat{C}_1, \ldots, \hat{C}_k$ are the resulting clusters and $\hat{\pi}_j \equiv \frac{|\hat{C}_j|}{n}$.

  *(i)* $\hat{\pi}_j \xrightarrow{\mathbb{P}} \pi_j$, for $j = 1, \ldots, k$.

  *(ii)* $\frac{\bar{D}}{\lambda_n} \xrightarrow{\mathbb{P}} 1$

  *(iii)* $\hat{W}_{ab} \equiv \frac{2\{\#\text{ of edges between } \hat{C}_a \text{ and } \hat{C}_b\}}{n\lambda_n} \xrightarrow{\mathbb{P}} W_{ab}$

- The algorithm does not perform well compared to spectral methods.

## *Other methods which work well for block models under regime (a)*

1. **Spectral Clustering:** Rohe et. al.(2011), McSherry (1993), Dasgupta, Hopcroft and McSherry (2004), Chaudhuri, Chung and Tsiatis (2012) develop methods which classify perfectly for $\frac{\lambda_n}{\log n} \to \infty$ (CCT (2012).
2. **Pseudo-likelihood** (with a good starting point). (Chen, Arash Amini, Levina and B (2012))
3. Methods based on empirical distribution of geodesic distances. (Bhattacharyya and B (2012))

# *Count statistics*

- Another approach works broadly even for $\lambda_n = O(1)$ regime.
- **Count statistics** are normalized subgraph counts and smooth functions of them.
- The subgraph count, $T(R)$, for subgraph corresponding to edge set $R$ is -

$$T(R) = \sum_{S \text{ subgraph of } G} \mathbf{1}(S \sim (V_R, R))/C(R, n)$$

where, $C(R, n) \equiv \#$ of all possible subgraphs of type $R$ of $G_n$.

- Example:
  - *(a)* Total number of triangles in a graph is an example of count statistic.
  - *(b)* Homophily $:= \frac{\#\text{ of }\Delta\text{'s}}{\#\text{ of }\Delta\text{'s} + \#\text{ of } V's}$ is an example of smooth function of count statistics.

# $O(1) \leq \lambda_n \leq O(n)$

Note that,

- $\rho_n \to 0,\ T(R) \xrightarrow{P} 0.$

But,

- If $|R| = p,\ \frac{\mathbb{E}(T(R))}{\hat{\rho}^p}$ stabilizes.

# *Theorem on Moment Estimate*

## *Theorem 1*

*I.* In dense case, if $p$, $R$ are fixed

$$\sqrt{n}\,(T(R) - \mathbb{E}(T(R))) \Rightarrow N(0, \sigma^2(R, P))$$

*II.* In general non-dense case conclusion of *I.* applies to $\frac{T(R)}{\hat{\rho}^p}$ where $|R| = p$ and $\hat{\rho} \equiv \frac{\bar{D}}{n}$, if $R$ is acyclic. Else result depends on $\lambda_n$.

# *The $\lambda_n = O(1)$ Regime*

- **Major difficulty**: For $\rho = \frac{\lambda_n}{n}$,

  $\mathbb{E}[\#$ of isolated points$] \sim ne^{-\lambda_n}$.

- We expect isolated points to correspond to special ranges of $\xi$.

  Example: For block models, $\pi_a$ small and $W_{ab}$ small for all $b$.

**Theorem (Decelle et. al. (2011)), (Mossel, Neeman and Sly (2012))**

For particular ranges of $(\pi, W, \lambda_n)$ block models can not be distinguished from Erdos-Renyi models. This includes but is not limited to $\lambda_n \leq 1$.

## *Consequence (Bickel, Chen and Levina (2011))*

For all $\lambda_n$ regimes, block models with fixed $K$ if $\mathbf{1}, Q\mathbf{1}, \ldots, Q^{k-1}\mathbf{1}$ are linearly independent with $Q_{ij} \equiv \frac{W_{ij}}{\pi_i}$,

- $\pi$ is estimable at rate $n^{-1/2}$.
- $S$ is estimable at rate $(n\lambda_n)^{-1/2}$
- $\bar{D} = \lambda_n(1 + O_{\mathbb{P}}((n\lambda_n)^{-1/2})$

(all $\lambda_n > 0$).

# *Identifiability*

The independence condition implies

1. $\mathbf{1}$ is not an eigenvector of $W$.
2. Block models with $\pi_1 = \cdots = \pi_K$ are not identifiable by acyclic count statistics.

# *A Computational Issue*

- Computing $T(R)$ for $R$ complex and $\sigma^2(T, R)$ is non-trivial.
- **Possible methods:** Estimates counts using
  - *(a)* Naive Monte Carlo (Sets of $m$ out of $n$ vertices).
  - *(b)* Weighted sampling of edges (Kashtan et. al. (2004)).
  - *(c)* Combination of above two methods (Bhattacharyya and B (2012)).

# *Discussion*

- Count statistics computation needs "bootstrap". (S. Bhattacharyya (2012))
- Simulations and real data applications available but much needed.
- **Block models:** In theory
  - $\frac{\lambda_n}{\log n} \to \infty$ and $\tau$ satisfying Property CAN are well-understood. $\frac{\lambda_n}{\log n} \to \infty$ is same as if the class memberships are known. Fast algorithms being developed in this regime.
  - $\lambda_n \to \infty$ but $\frac{\lambda_n}{\log n} \not\to \infty$ consistent but not $\sqrt{n}$-consistent estimation possible
  - $\lambda_n = O(1)$ very subtle.
- Many statistical extensions needed: covariates, dynamics etc.
- If $\lambda_n \to \infty$ should be able to formulate minimax result for estimation of $w(\xi_1, \xi_2)$ or more generally $w(\mathbf{z}_1, \mathbf{z}_2)$ for "sparse" situations under smoothness conditions on $w(\cdot, \cdot)$.