Community detection in networks - a probabilistic approach

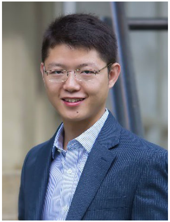
Anirban Bhattacharya

February 17, 2017
Texas A&M University, College Station

Acknowledgements: Collaborators

Junxian Geng
FSU

Debdeep Pati
FSU



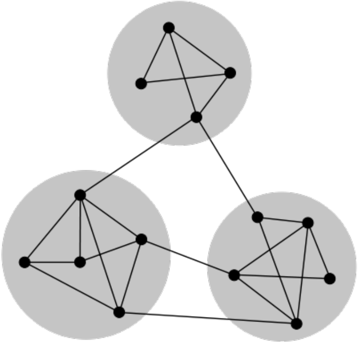
Zhengwu Zhang
SAMSI

Outline of the Talk

- Motivation
- Clustering $\sim$ community detection in networks
- Literature review
- MFM-SBM
- Numerical Illustrations
- Marginal likelihood analysis
- Applications on brain connectivity network
- Ongoing work

Motivation

- social networks, connectomics, biological networks, gene circuits, internet networks (Goldenberg, Zheng, Fienberg & Airoldi, 2010)
- One typical sparsity pattern: groups of nodes with dense within group connections and sparser connections between groups.



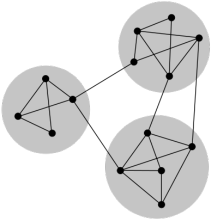
Mathematical Formulation

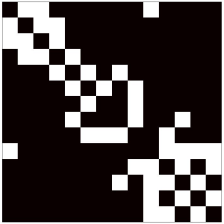
- Observable: $G = (V, E)$ undirected / directed graph
- $V = \{1, 2, \ldots, n\}$ arbitrarily labelled vertices
- $A(n \times n)$ adjacency matrix encoding edge information
-

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge (relationship) between (from) } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

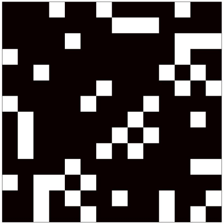
- We assume $A_{ii} = 0$ (but self loops can be allowed)

Adjacency Matrix (undirected)





Nodes in order



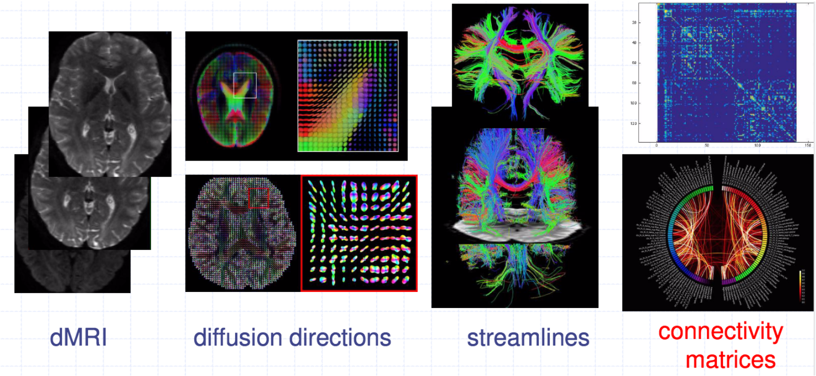
random order

Community detection

- Goal:
  1. Learn the number of communities (k) and
  2. Cluster the nodes which share a similar connectivity pattern

Biological networks: Human connectomics data

- Diffusion Tensor Imaging (DTI) provides a reliable connectivity measure.
- An illustration of a standard pipeline (Hagmann, 2005) of extracting diffusion MRI (dMRI) to connectomics data.
- Goal: Cluster the 68 brain regions (34: LH, 34: RH) based on connections.



Existing methods for community detection

- Large literature on community detection in networks
- Graph-theoretic, Modularity, Spectral, Maximum likelihood and Bayesian
- Nowicki & Snijders (2001), Newman & Girvan (2004), Zhao, Levina & Zhu (2011), Rohe, Chatterjee & Wu (2011), Chen, Bickel & Levina (2013), Abbe & Sandon (2015) . . .

Existing methods for community detection

- Assume knowledge of the number of communities (Airoldi et al., 2009; Bickel and Chen, 2009; Amini et al., 2013) or estimate it apriori using either of cross-validation, hypothesis testing, BIC or spectral methods (Daublin et al., 2008; Latouche et al., 2012; Wang and Bickel, 2015; Lei, 2014; Chen & Lei, 2014; Le and Levina, 2015)
- 2-stage procedures ignore uncertainty in the first stage and are prone to increased misclassification
- Existing Bayesian methods for unknown k: both conceptual and computational issues.
- Our goal is to propose a coherent probabilistic framework with efficient sampling algorithms which allows simultaneous estimation of the number of clusters and the cluster configuration.

Stochastic Block Model (Holland et al., 1983)

- A parsimonious model favoring block structure
- $A_{ij} \sim \text{Bernoulli}(\theta_{ij})$, with θ_{ij} characterized by community memberships
- Nodes belong to one of k communities, let $z_i \in \{1, \ldots, k\}$ denote the community membership of the ith node
- $Q = (Q_{rs}) \in [0, 1]^{k \times k}$, with Q_{rs} the probability of an edge from any node i in cluster r to any node j in cluster s
- $$A_{ij} \sim \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}$$
- Assume $P(z_i = j) = \pi_j, \quad j = 1, \ldots, k$

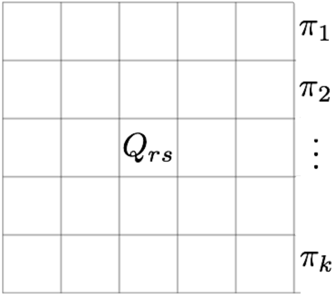
$$P(A_{ij} = 1) = \sum_{r=1}^{k} \sum_{s=1}^{k} Q_{rs} \pi_r \pi_s = \pi^{\mathrm{T}} Q \pi$$

Generalization to Random graph models

- Under node exchangeability assumption Aldous & Hoover, 1981 showed that there exists $\xi_i \sim \mathrm{U}(0,1)$ and a graphon $h : [0,1] \times [0,1] \to [0,1]$ such that

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

- SBM: h is constant $Q_{r,s}$ on block (r,s) of size $\pi_r \times \pi_s$.



Graphon of SBM

Bayesian formulation

- General framework for prior specification: With $z = (z_1, \ldots, z_n)$

$$
\begin{aligned}
&(z, k) \sim \Pi \\
&Q_{rs} \overset{\text{ind}}{\sim} U(0,1), \quad (r, s = 1, \ldots, k), \\
&A_{ij} \mid z, Q, k \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}
\end{aligned}
$$

- Π is a probability distribution on the space of partitions of $\{1, \ldots, n\}$
- Nowicki and Snijders (2001): Assumes known k, and

$$
\begin{aligned}
z_i \mid \pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_k) \\
\pi &\sim \text{Dir}(\alpha/k, \ldots, \alpha/k)
\end{aligned}
$$

- Carvalho et al 2015: Assumes unknown k through Chinese restaurant process.

Carvalho et al 2015: CRP based prior for (z, k)

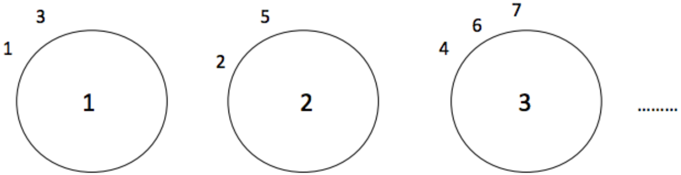
- A possible model for z_i:

$$z_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_k)$$
$$\pi \sim \text{Dir}(\alpha/k, \ldots, \alpha/k)$$

- As $k \to \infty$, Ishwaran and Zarepour (2002) showed that the distribution of z_is:

$$p(z_i = c \mid z_{-i}) \sim \begin{cases} |c| & \text{at an existing table c} \\ \alpha & \text{if c is a new table} \end{cases}$$

where $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$



Some discussion on CRP

- Partitions sampled from the CRP posterior tend to have multiple small transient clusters.
- Let t be the number of clusters (tables) $s = (s_1, \dots, s_t)$ denotes the vector of cluster sizes, then

$$\boxed{P(S = s) = V_n^{CRP}(t) \frac{n!}{t!} s_1^{-1} \dots s_t^{-1}}$$

- Probability of small, transient clusters high
- inconsistent estimation of the number of clusters (Miller and Harrison, 2015)

Mixture of finite mixtures (MFM)

Mixture of finite mixture (MFM) model (Miller & Harrison, 2016+):

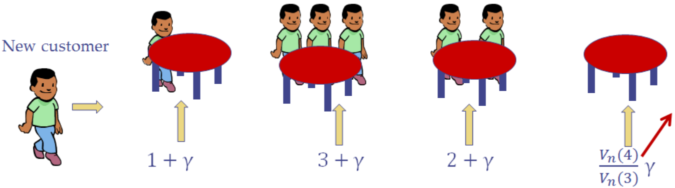
$$z_i \mid \pi, k \sim \text{Multinomial}(\pi_1, \ldots, \pi_k)$$
$$\pi \mid k \sim \text{Dir}(\gamma, ..., \gamma)$$
$$k \sim p(\cdot), \text{where } p(\cdot) \text{ is a proper p.m.f on 1,2,...}$$

$$\boxed{P(S = s) = \frac{V_n(t)}{\Gamma(\gamma)^t} \frac{n!}{t!} \prod_{i=1}^{t} s_i^{\gamma-1}}$$

Modified Chinese restaurant process (m-CRP)

$$p(z_i = c \mid z_{-i}) \sim \begin{cases} |c| + \gamma & \text{at an existing table c} \\ \dfrac{V_n(t+1)}{V_n(t)}\gamma & \text{if c is a new table} \end{cases}$$

$V_n(t+1)$: pre-stored sequences



γ controls the probability of introducing new clusters.

Complete prior specification (MFM-SBM)

- The model along with the prior specified above can be expressed hierarchically as follows:

$$\boxed{k \sim p(\cdot), \text{where } p(\cdot) \text{ is truncated Poisson } \{1, \ldots, \text{n}\}}$$

$$Q_{rs} \overset{\text{ind}}{\sim} \text{Unif}(0,1), \quad r, s = 1, \ldots, k,$$

$$\boxed{\pi \sim \text{Dirichlet}(\gamma, \ldots, \gamma)},$$

$$P(z_i = j \mid \pi) = \pi_j, \quad i = 1, \ldots, n; j = 1, \ldots, k$$

$$A_{ij} \mid z, Q \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}.$$

MCMC algorithm

- Marginalization of k possible due to modified CRP
- No need to perform RJMCMC / allocation samplers
- Efficient Gibbs sampler updates for z and Q

Data Generation

- Decide the number of communities k and the number of subjects n.
- Set the true cluster configuration of the data $z_0 = (z_{01}, \dots, z_{0n}) : z_{0i} \in \{1, \dots, k\}$.
- Set values for edge probability matrix $Q = (Q_{rs}) \in [0,1]^{k \times k}$.

$$Q = \begin{bmatrix} p & 0.1 & \dots & 0.1 \\ 0.1 & p & \dots & 0.1 \\ \vdots & \vdots & \ddots & \vdots \\ 0.1 & 0.1 & \dots & p \end{bmatrix}$$

The smaller p is, the more vague the block structure

- Finally, generate the adjacency matrix A from $\text{Bernoulli}(Q_{z_i z_j})$.
- Use Rand Index $(\#$ of "agreement pairs" $/ \binom{n}{2})$ to compare estimation of z

Comparison with existing methods

- Hyperparameters: use $\gamma = 1$, truncated Poisson(1)
- Investigate mixing and convergence vs. CRP-SBM.
- Compare estimation of both z and k

Mixing / convergence comparison

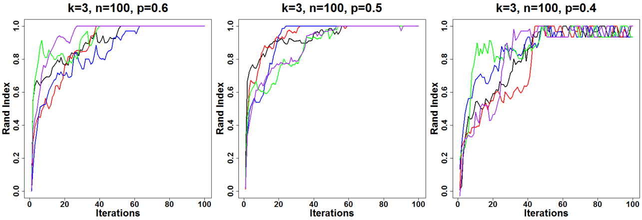


Figure: MFM-SBM, balanced network, 100 nodes in 3 communities

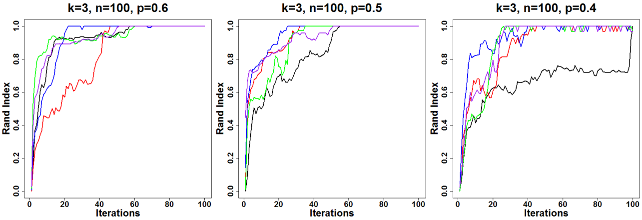


Figure: MFM-SBM, unbalanced network, 100 nodes in 3 communities.

Mixing / convergence comparison

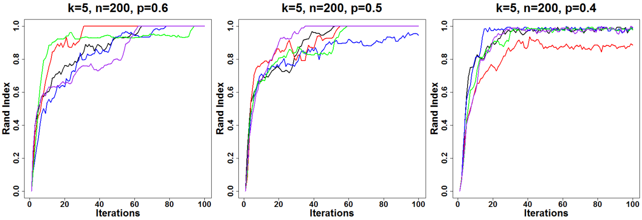


Figure: MFM-SBM, unbalanced network, 200 nodes in 5 communities.

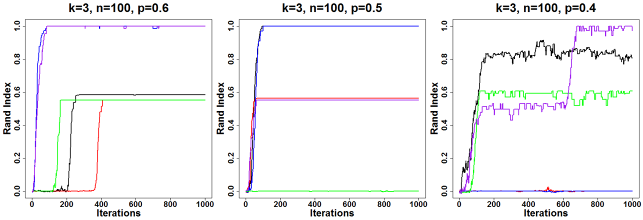


Figure: CRP-SBM, balanced network, 100 nodes in 3 communities.

Comparison on estimating (k, z)

- Two settings:
  1. Well-specified setting: $\theta_{ij} = Q_{z_i,z_j}$
  2. Misspecified setting: $\theta_{ij} = w_i w_j Q_{z_i,z_j}$, 30% of the nodes have $w_i = 0.7$, remaining $w_i = 1$.
- (k, z) estimated using Zhang, Pati & Srivastava, 2015.
- Comparison based on the $N = 100$ replicated datasets
- Competitors based on spectral properties of certain graph operators, namely the i) non-backtracking matrix (NBM) ii) Bethe Hessian matrix (BHM) → Le & Levina, 2016 iii) Leading eigen vector method (LEM) Newman, 2006 iv) Hierarchical modularity measure (HMM) Blondel et al 2008 v) B-SBM (allocation based sampler version of our method)

Specified case: Comparison on estimating k

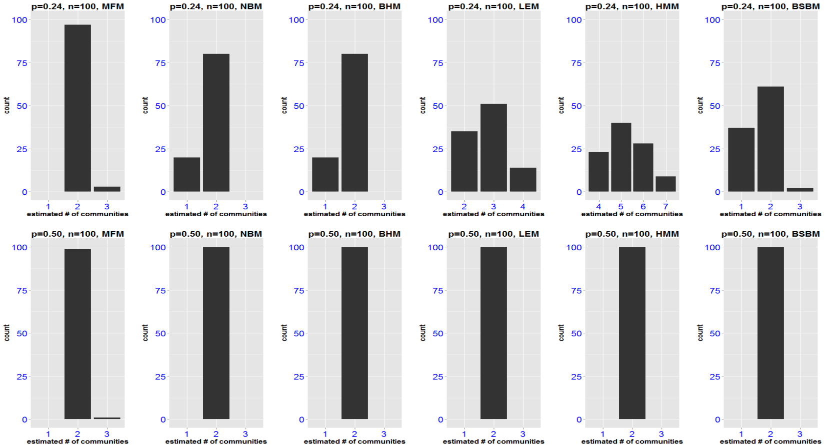


Figure: 2 communities and same size, left to right: our method, competitor I, competitor II

Specified case: Comparison on (z, k) estimation

(k, p)	**MFM-SBM**	**LEM**	**HMM**	**B-SBM**
$k = 2, p = 0.50$	0.99 (1.00)	1.00 (0.99)	1.00 (1.00)	1.00 (1.00)
$k = 2, p = 0.24$	0.97 (0.84)	0.35 (0.79)	NA (NA)	0.61 (0.78)
$k = 3, p = 0.50$	1.00 (1.00)	0.67 (0.96)	1.00 (0.99)	0.91 (0.99)
$k = 3, p = 0.33$	0.97 (0.93)	0.85 (0.79)	0.78 (0.89)	0.54 (0.93)

Table: The value outside the parenthesis denotes the proportion of correct estimation of the number of clusters out of 100 replicates. The value inside the parenthesis denotes the average Rand index value when the estimated number of clusters is correct.

Misspecified case: Comparison on estimating k

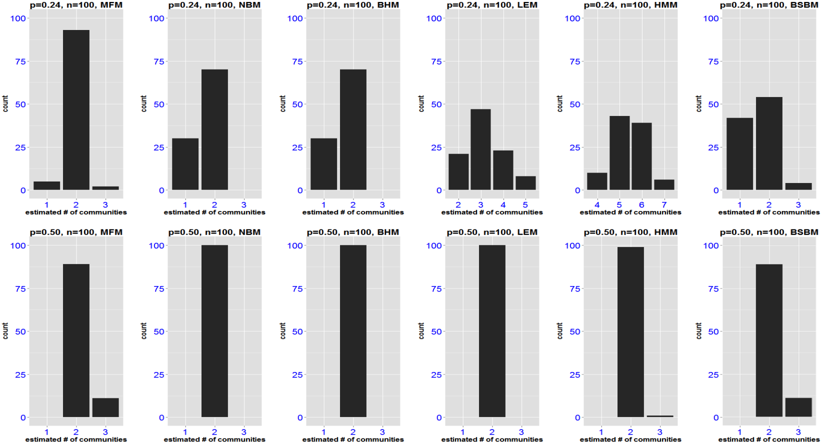


Figure: 2 communities and same size, left to right: our method, competitor I, competitor II

Misspecified case: Comparison on (z, k) estimation

(k, p)	**MFM-SBM**	**LEM**	**HMM**	**B-SBM**
$k = 2, p = 0.50$	0.90 (1.00)	1.00 (1.00)	0.99 (1.00)	0.89 (1.00)
$k = 2, p = 0.24$	0.93 (0.80)	0.21 (0.73)	NA (NA)	0.54 (0.57)
$k = 3, p = 0.50$	0.96 (0.99)	0.75 (0.94)	1.00 (0.99)	0.87 (0.99)
$k = 3, p = 0.33$	0.93 (0.88)	0.78 (0.73)	0.47 (0.80)	0.38 (0.82)

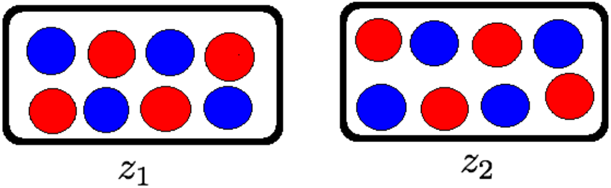
Table: The value outside the parenthesis denotes the proportion of correct estimation of the number of clusters out of 100 replicates. The value inside the parenthesis denotes the average Rand index value when the estimated number of clusters is true.

Clustering consistency

- Very rich literature in Bayes theory on parameter estimation
- Considerably smaller literature on estimation of discrete configurations (Johnson & Rossell, 2012, Narisetty & He, 2014; etc)
- No general results on $\boxed{\text{clustering consistency}}$ in Bayes paradigm
- There are a few frequentist results on consistent community detection (Abbe & Sandon (2015; 2016), Bickel and Chen, 2009; Dembo et al 2015, Mossel et al 2014) under the planted partition model
- Generally assumes Q and k_0 are known

Mode of convergence to the true community assignment

- Community assignments identifiable up to arbitrary labeling of the community indicator within each community.



$$d(z_1, z_2) = 0$$

- Consider *d*: minimum Hamming distance on *equivalence classes* of community assignments which are identical modulo labeling / *n*

Clustering consistency for sparse networks

- (A1) Assume $k_0 = 2$ with roughly equal sized communities.
- (A2) Network is homogeneous, i.e.

$$Q^0 = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

with $0 \leq q < p \leq 1$.

Clustering consistency: Main results

Define $\mathrm{I} = D(p||q)$, the Renyi divergence of order $\frac{1}{2}$ between p and q.

Theorem

Under (A1) and (A2), model fitted with $k = 2$ *&* $\gamma = 1$

$$\mathbb{E}[d(z, z_0) \mid \mathcal{A}] \leq \exp\left[-\{1 + o(1)\}\frac{n\mathrm{I}}{2}\right] \tag{1}$$

- This is the optimal rate of convergence Zhang & Zhou, 2016+.

Theorem

Under (A1) and (A2), model fitted with a Poisson prior on k

$$\mathbb{E}[d(z, z_0) \mid \mathcal{A}] \to 0 \quad as \quad n \to \infty. \tag{2}$$

Interpretation of the result: Sparse networks

- Note that if $p = \frac{a}{n}$ and $q = \frac{b}{n}$

$$n\mathrm{I} = nD\left(\frac{a}{n}\middle\|\frac{b}{n}\right) \asymp n\frac{(a-b)^2}{an} = \frac{(a-b)^2}{a}.$$

- When k_0 is known, consistent community detection when $\frac{(a-b)^2}{a} \to \infty$.
- If $n\mathrm{I} = 2\rho \log n$, then we roughly misclassify $n^{1-\rho}$ nodes.
- If $\rho > 1$, exact recovery is possible asymptotically.
- Partial recovery for $\rho \leq 1$.

Proof technique overview

We look at the posterior expected loss respect to distance d on

$$\begin{aligned}\mathbb{E}[d(z, z_0) \mid \mathcal{A}] &= \sum_r r \frac{\sum_{z:d(z,z_0)=r} \exp\{\ell(z \mid \mathcal{A}) - \ell(z_0 \mid \mathcal{A})\}\{\pi(z)/\pi(z_0)\}}{\sum_z \exp\{\ell(z \mid \mathcal{A}) - \ell(z_0 \mid \mathcal{A})\}\{\pi(z)/\pi(z_0)\}} \\ &\leq \sum_r r \sum_{z:d(z,z_0)=r} \exp\{\ell(z \mid \mathcal{A}) - \ell(z_0 \mid \mathcal{A})\}\{\pi(z)/\pi(z_0)\}.\end{aligned}$$

- small r: Difficult to "separate" log-marginal likelihood (LML) for z and z_0, but $|z : d(z, z_0) = r|$ is small (low model complexity)
- larger r: Easy separation between LMLs, but higher model complexity

Separation analysis of LMLs (fixed k)

- For any z, z^1, z^2, let

$$\begin{aligned} n_{r,s}(z) &= \sum_{i<j} \mathbb{1}(z_i = r, z_j = s), \\ a_{r,s}(z) &= \sum_{i<j} a_{ij} \mathbb{1}(z_i = r, z_j = s) \\ n_{r,s;l,t}(z^1, z^2) &= \sum_{i<j} \mathbb{1}(z_i^1 = r, z_j^1 = s) \mathbb{1}(z_i^2 = l, z_j^2 = t) \end{aligned}$$

  for $1 \leq r, s, l, t \leq k$.

- log marginal likelihood: for $h(x) = x \log x + (1 - x) \log(1 - x)$

$$\ell(\mathcal{A} \mid z) = \boxed{\sum_{r,s} n_{r,s}(z) h\left\{ \frac{a_{r,s}(z)}{n_{r,s}(z)} \right\}} - \log \sum_{r,s} \{n_{r,s}(z) + 1\}$$

- Boxed part: $\tilde{\ell}(\mathcal{A} \mid z)$

Separation analysis of LMLs (fixed k)

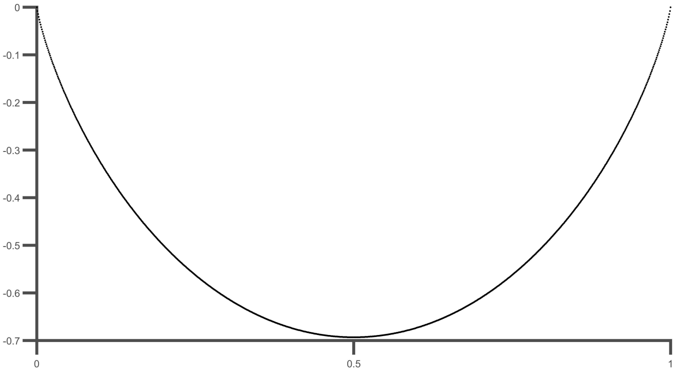


Figure: Plot of $h(x) = x \log x + (1-x) \log(1-x)$ for $x \in [0,1]$. h is symmetric about $1/2$ where it attains its minima on $[0,1]$.

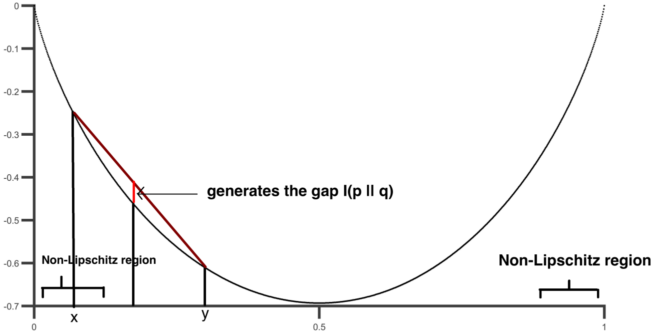
Separation analysis of LMLs (fixed k)

$$\begin{aligned}\ell(\mathcal{A} \mid z_0) - \ell(\mathcal{A} \mid z) \;&\asymp\; \underbrace{\left\{\tilde{\ell}(\mathcal{A} \mid z_0) - \tilde{\ell}_E(\mathcal{A} \mid z_0)\right\} - \left\{\tilde{\ell}(\mathcal{A} \mid z) - \tilde{\ell}_E(\mathcal{A} \mid z)\right\}}_{\text{2 stochastic terms}} \\ &+\; \underbrace{\left\{\tilde{\ell}_E(\mathcal{A} \mid z_0) - \tilde{\ell}_E(\mathcal{A} \mid z)\right\}}_{\text{non-stochastic term}}\end{aligned}$$

- $a_{r,s;l,t}(z, z_0) \sim \mathrm{Bernoulli}(n_{r,s;l,t}(z, z_0), Q^0_{lt})$, consider the "population" versions of $\tilde{\ell}(\mathcal{A} \mid z)$ and $\tilde{\ell}(\mathcal{A} \mid z_0)$ as

$$\begin{aligned}\tilde{\ell}_E(\mathcal{A} \mid z) &= \sum_{r,s;l,t} n_{r,s;l,t}(z, z_0) h\left\{\frac{\sum_{l,t} n_{r,s;l,t}(z, z_0) Q^0_{lt}}{\sum_{l,t} n_{r,s;l,t}(z, z_0)}\right\}, \\ \tilde{\ell}_E(\mathcal{A} \mid z_0) &= \sum_{r,s;l,t} n_{r,s;l,t}(z, z_0) h(Q^0_{lt}).\end{aligned}$$

Separation analysis of LMLs (fixed k)



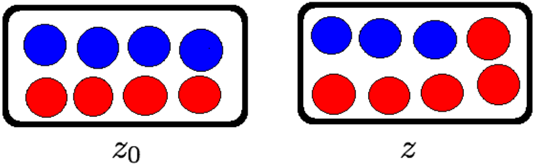
- Non-stochastic term: Since h is a convex function,

$$h\left\{\frac{\sum_{l,t} n_{r,s;l,t}(z,z_0)Q^0_{lt}}{\sum_{l,t} n_{r,s;l,t}(z,z_0)}\right\} < \frac{1}{\sum_{l,t} n_{r,s;l,t}(z,z_0)} \sum_{l,t} n_{r,s;l,t}(z,z_0)h(Q^0_{lt})$$

- Stochastic term: Since h is non-Lipschitz near boundary, we use stronger non-linear large-deviation theory

Separation analysis of LMLs (fixed k)

- An example:



$$d(z_0, z) = 1$$

-

$$\tilde{\ell}_E(z \mid \mathcal{A}) - \tilde{\ell}_E(z_0 \mid \mathcal{A}) \leq -Cn \times \mathrm{I}(p||q)$$

Separation analysis of LMLs (fixed k)

- For fixed k_0, when $d(z, z_0) = r$,

  $$\tilde{\ell}_E(z \mid \mathcal{A}) - \tilde{\ell}_E(z_0 \mid \mathcal{A}) \leq -C_1 nr \times \mathrm{I}(p||q)$$

  with high probability.
- Also for fixed k_0

  $$\max_{z : d(z, z_0) \leq r} \left\{ \log \pi(z) - \log \pi(z_0) \right\} \leq C_2 r$$

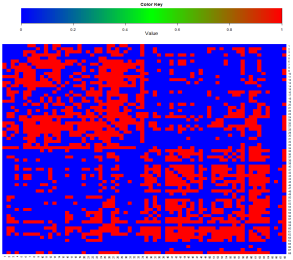
- The above two observations conclude the proof.

Real data examples

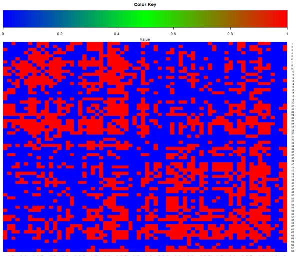
Human connectomics data

- Goal: Cluster the 68 regions (34: LH, 34: RH) based on connections
- Biological importance of the clusters: Brain activation regions (Smith et al 2015)
- Connections between left and right hemisphere are believed to have positive impact on cognitive ability
- Analyze connectomes of two subjects with different inter-hemisphere connections

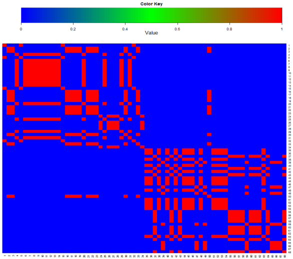
Subjects 1 and 8



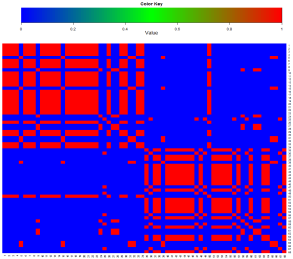
Adjacency Matrix (Sub 1)



Adjacency Matrix (Sub 8)



$\hat{B}_{ij} = \mathbb{1}(\hat{z}_i = \hat{z}_j)$: MFM-SBM
(3-LH+ 3-RH)



$\hat{B}$: MFM-SBM
(3-LH+3RH+1Mixed)

Conclusion

- For subjects 1 and 8, clusters obtained from MFM-SBM in LH/ RH involve
  1. Sensory and motor regions
  2. insula and dorsal attention regions
  3. cerebellar regions
- Regions conform with Smith et al 2015 Nature Neuroscience paper using FMRIB Software Library
- For subject 8, one cluster has nodes from insular region of both LH and RH
- B-SBM performs closely, LEM & HMM favored smaller clusters

Dolphin social network data: Reference configuration

- 62 dolphins living off Doubtful sound, NZ from 1994-2001

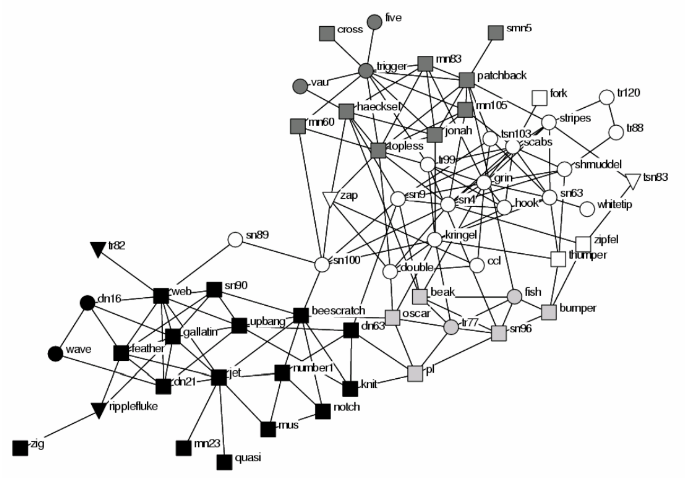


Figure: Reference configuration obtained from Lusseau et al 2003

Dolphin social network data: MFM-SBM

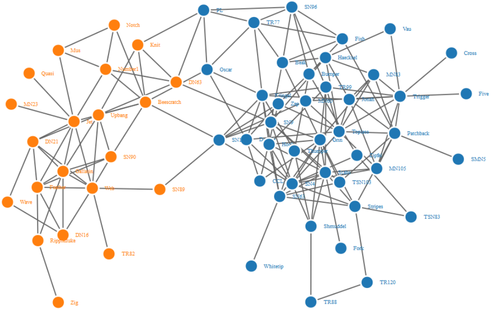
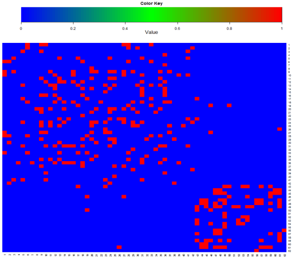


Figure: Perfect clustering except one subject

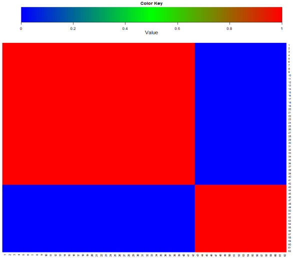
MFM-SBM	NBM	BHM	LEM	HMM	B-SBM
2	2	2	5	5	3

Table: Estimated number of clusters for dolphin data

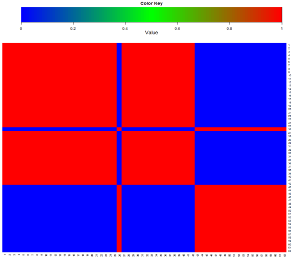
Dolphin social network data: Comparison



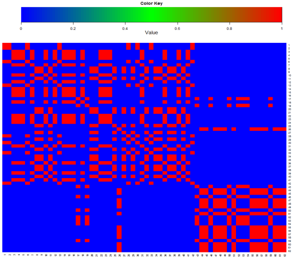
Adjacency Matrix



$\hat{B}_{ij} = \mathbb{1}(z_i^0 = z_j^0)$: Reference



$\hat{B}$: MFM-SBM



$\hat{B}$: HMM

Discussion & Ongoing directions on connectomics data

- MFM-SBM has good performance in simulated and real-data examples.
- Model for population of connectomes
- Allow subject specific deviations, adjust for covariates
- Can potentially fit increasingly complex models
- Model adequacy checks very important

Acknowledgements: Grants

NSF DMS-1613156 (PI)

ONRBAA14-001 (PI)

Reference

- Probabilistic community detection with unknown number of communities, http://arxiv.org/abs/1602.08062 with Anirban Bhattacharya and Junxian Geng, *under revision at the Journal of the American Statistical Association*.

Thank You !

Demonstrating uncertainty

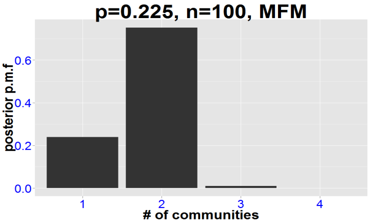


Figure: true $k = 2$

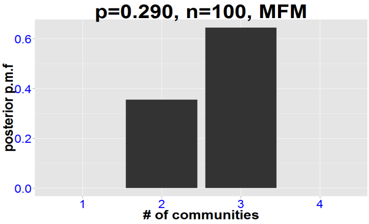
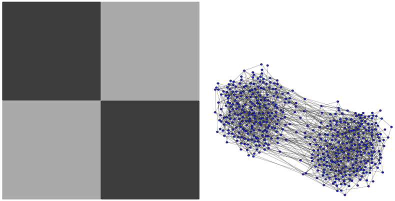


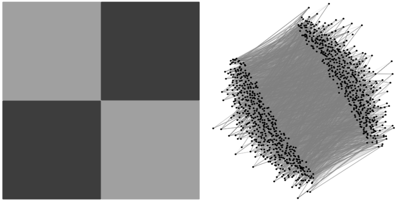
Figure: true $k = 3$

- substantial uncertainty when the community structure is not prominent
- MFM-SBM of avoids tiny extraneous communities as opposed to CRP type models.

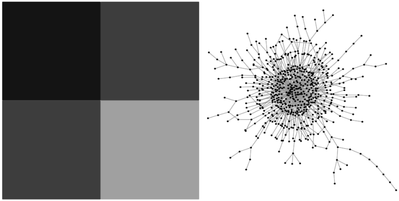
Different types of SBMs



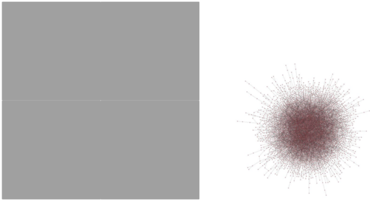
(a) Community structure



(b) Disassortative structure



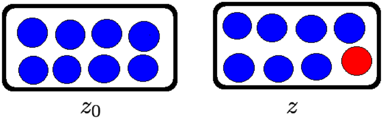
(c) Core-periphery structure



(d) Random graph

Strength in Marginal likelihood ($k_0 = 1$ vs $k = 2$)

- $\Pi(k = k_0 \mid \mathcal{A}) \to 1$ due to Allman et al 2009.
- An example:



$d(z_0, z) = 1$

- log marginal likelihood: for $h(x) = x \log x + (1-x)\log(1-x)$

$$\ell(\mathcal{A} \mid z) = \sum_{r,s} n_{r,s}(z) h\left\{\frac{a_{r,s}(z)}{n_{r,s}(z)}\right\} - \boxed{\log \sum_{r,s}\{n_{r,s}(z)+1\}}$$

- $\boxed{\cdot}$ comes to the rescue ! here as $\tilde{\ell}_E(z \mid \mathcal{A}) = \tilde{\ell}_E(z_0 \mid \mathcal{A})$.
- $\exp\{\boxed{\cdot}\} = 1/n^3$ for z and $1/n^2$ for z_0.