Dependence Structures in Network Data

Sharmodeep Bhattacharyya

Oregon State University

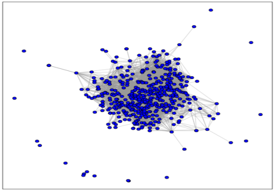
Statistics in Big Data Era, June 2022

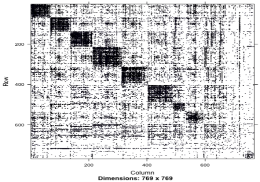
Main Collaborators

Figure: Shirshendu Chatterjee, CUNY and Soumendu Sundar Mukherjee, ISI, Kolkata

Outline

Network Data



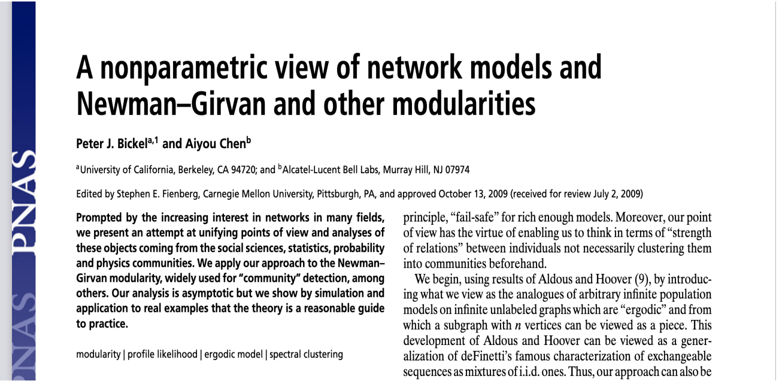


- Adjacency matrices (Symmetric), $[A_{ij}]_{i,j=1}^{n}$ numerically represent network data:

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ links to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

- $\mathbf{A_{ij}}$ are **dependent** on each other.
  Question: How to model this dependence?

Exchangeable Network Models



PNAS

A nonparametric view of network models and Newman–Girvan and other modularities

Peter J. Bickel[a,1] and Aiyou Chen[b]

[a]University of California, Berkeley, CA 94720; and [b]Alcatel-Lucent Bell Labs, Murray Hill, NJ 07974

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved October 13, 2009 (received for review July 2, 2009)

Prompted by the increasing interest in networks in many fields, we present an attempt at unifying points of view and analyses of these objects coming from the social sciences, statistics, probability and physics communities. We apply our approach to the Newman–Girvan modularity, widely used for "community" detection, among others. Our analysis is asymptotic but we show by simulation and application to real examples that the theory is a reasonable guide to practice.

modularity | profile likelihood | ergodic model | spectral clustering

principle, "fail-safe" for rich enough models. Moreover, our point of view has the virtue of enabling us to think in terms of "strength of relations" between individuals not necessarily clustering them into communities beforehand.

We begin, using results of Aldous and Hoover (9), by introducing what we view as the analogues of arbitrary infinite population models on infinite unlabeled graphs which are "ergodic" and from which a subgraph with n vertices can be viewed as a piece. This development of Aldous and Hoover can be viewed as a generalization of deFinetti's famous characterization of exchangeable sequences as mixtures of i.i.d. ones. Thus, our approach can also be

Figure: Peter and Aiyou's Seminal Work.

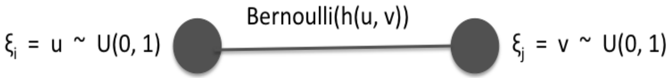
Exchangeable Network or Graphon Models

Derived from representation of exchangeable random infinite array by Aldous and Hoover (1983).

Exchangeable Network Model

Define $\mathbb{P}(\{A_{ij}\}_{i,j=1}^{n})$ conditionally given latent variables $\{\xi_i\}_{i=1}^{n}$ associated with vertices $\{v_i\}_{i=1}^{n}$ respectively. (Bickel & Chen (2009), Bollobás et.al. (2007), Hoff et.al. (2002)).

$$
\begin{aligned}
\xi_1, \ldots, \xi_n &\overset{\text{iid}}{\sim} \mathcal{U}(0,1) \\
\mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v) &= h_n(u,v) = \rho_n w(u,v),
\end{aligned}
$$



Sequence of Networks

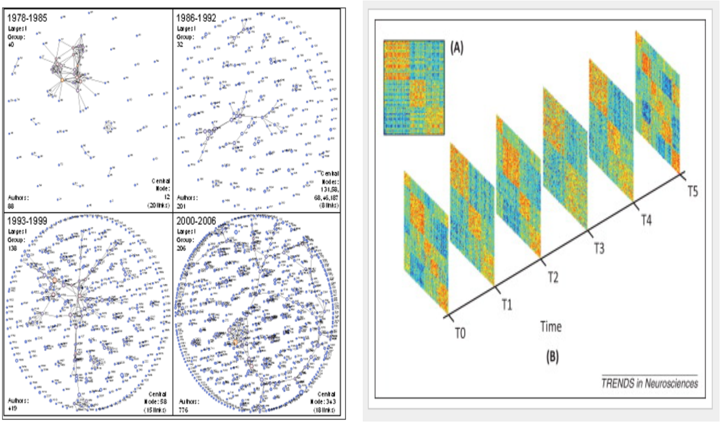


Figure: Network Sequence Examples

Voting Patterns in US Congress

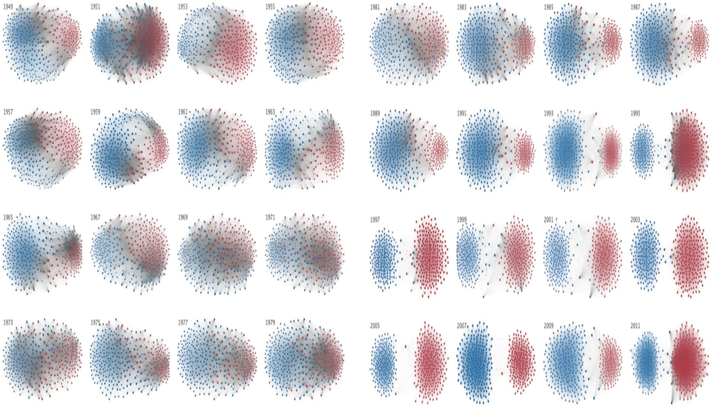
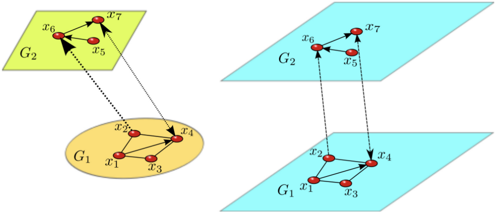


Figure: Network Sequence Examples

Multi-layer Network Data



Definition (Multi-layer network)

A *multi-layer network* is a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ where $\mathcal{G} = \{G_\alpha; \alpha \in \{1, \ldots, M\}\}$ is a family of (directed or undirected, weighted or unweighted) graphs $G_\alpha = (X_\alpha, E_\alpha)$ (called layers of $\mathcal{M}$) and $\mathcal{C} = \{E_{\alpha\beta} \subseteq X_\alpha \times X_\beta; \alpha, \beta \in \{1, \ldots, M\}, \alpha \neq \beta\}$ is the set of interconnections between nodes of different layers G_α and G_β with $\alpha \neq \beta$.

Multi-layer Network Data

- A **multiplex network** is a special type of multilayer network in which $X_1 = X_2 = \cdots = X_M = X$ and the only possible type of interlayer connections are between corresponding nodes in different layers, that is, $E_{\alpha\beta} = \{(x, x); x \in X\}$ for every $\alpha, \beta \in \{1, \ldots, M\}, \alpha \neq \beta$.
- A **temporal network** $(G(t))_{t=1}^{T}$ can be represented as a multilayer network with a set of layers $\{G_1, \ldots, G_T\}$ where $G_t = G(t)$, $E_{\alpha\beta} = \emptyset$ if $\beta \neq \alpha + 1$, while $E_{\alpha,\alpha+1} = \{(x, x); x \in X_\alpha \cap X_{\alpha+1}\}$. Notice that here t is an integer, and not a continuous parameter.

Statistical Inference for Dependent Networks

1. **Network Sequences with dependent layers:** We consider network sequences with two different types of dependent layers.
    - **Dependence of network layer adjacency matrices:** The adjacency matrices corresponding to each of the network layers are dependent over of the sequence.
    - **Dependence of network layer model parameters:** The model parameters or latent distributions of each network layer models are dependent over the sequence.
2. **Networks with dependent edge formation:** We consider sequences of networks with dependence edge construction mechanism, going beyond the *graphon* models.

Main Questions

1. **Community Detection:** Estimation of latent memberships of the nodes using the edge structure of the single and multi-layer networks.
2. **Change-point Detection:** Estimation of structural break points in sequence of networks.
3. **Parameter Estimation:** Estimation of population parameters of the network models with edge-dependence and dependence between layers.

Outline

Two Forms of Inter-layer Dependencies

We consider network sequences with two different types of dependent layers.

- **Dependence of network layer adjacency matrices:** The adjacency matrices corresponding to each of the network layers are dependent over of the sequence.
- **Dependence of network layer model parameters:** The model parameters or latent distributions of each network layer models are dependent over the sequence.

Outline

Additional Collaborators

- Sayak Chatterjee, Indian Statistical Institute, Kolkata.
- Anirban Nath, Indian Statistical Institute, Kolkata.

Lazy inhomogeneous Erdős-Rényi graph

A sequence of $T \geq 2$ adjacency matrices $(A^{(1)}, \ldots, A^{(T)})$ is said to be generated from the lazy inhomogeneous Erdős-Rényi (lazy IER in abbreviation) process with parameters $\mathbf{P} \in [0,1]^{n \times n}$ satisfying $P_{ij} = P_{ji}$ for all $i, j \in [n]$ and $\alpha \in (0,1)$, if

$$A_{ij}^{(1)} \sim Bernoulli(P_{ij}), \text{ and}$$

$$A_{ij}^{(t)} = \begin{cases} A_{ij}^{(t-1)} & \text{with probability } \alpha \\ \overset{ind}{\sim} Bernoulli(P_{ij}) & \text{with probability } 1-\alpha \end{cases}$$

Lazy inhomogeneous Erdős-Rényi graph

- **Aggregated adjacency matrix** is defined as

$$A := \sum_{t \in [T]} A^{(t)}.$$

- **Aggregated Laplacian matrix** is defined as

$$\mathcal{L} := I_n - D^{-1/2} A D^{-1/2},$$

  where, D is the diagonal matrix of aggregated degrees $d_i := \sum_{t \in [T], j \in [n]} A_{ij}^{(t)}, i = 1, \ldots, n$.

- The **population** versions of the aggregated adjacency and Laplacian matrices are defined respectively as $\bar{A} := \mathbb{E}(A)$ and $\bar{\mathcal{L}} := I_n - \bar{D}^{-1/2} \bar{A} \bar{D}^{-1/2}$, where $\bar{D} = \mathbb{E} D$.

Concentration Around Population Parameter

Theorem (Chatterjee[2], Mukheree, Nath and B. (2022))

Let $d_{\min} := \min_{i\in[n]} \sum_{j\in[n]} P_{ij}$ *(resp.* $d_{\max} := \max_{i\in[n]} \sum_{j\in[n]} P_{ij}$*) denote the minimum (resp. maximum) among the expected degrees of the vertices for each* $t \in [T]$*. Then there exist constants* $C, C_1(\alpha) > 0$ *such that if* $Td_{\max} > C(\log(n))^3$*, then*

$$\|A - \bar{A}\| \leq C_1(\alpha)\sqrt{Td_{\max}\log(n)}$$

with high probability. Moreover, there exist constants $C, C_2(\alpha) > 0$ *such that if* $Td_{\min} > C(\log(n))^3$*, then*

$$\|\mathcal{L} - \bar{\mathcal{L}}\| \leq C_2(\alpha)\sqrt{\frac{\log(n)}{Td_{\min}}}$$

with high probability.

Concentration Around Population Parameter

- Both $C_1(\alpha), C_2(\alpha) \uparrow \infty$ as $\alpha \uparrow 1$, and $C_1(\alpha), C_2(\alpha) \downarrow C_0$ for an universal constant $C_0 > 0$ as $\alpha \downarrow 0$. In fact, both of these constants are $\mathbf{\Omega(1/\sqrt{1-\alpha})}$.
- For $\alpha = 0$, concentration is **highest**.
  For, α increasing to 1, the correlation among the edges between each pair of vertices across all layers increases to 1. Consequently, the concentration properties of the aggregated adjacency and Laplacian matrices deteriorate.
- Matrix Bernstein-type inequalities, combinatorial arguments, and path counting arguments used in random matrix theory are not useful for proving the nontrivial concentration results for strongly correlated layers of multi-layer network models that we address in Theorem. Our approach enables us to go beyond weakly correlated multi-layer network models and develop tools to analyze multi-layer network models involving more complex correlation structures.

Simulation Results

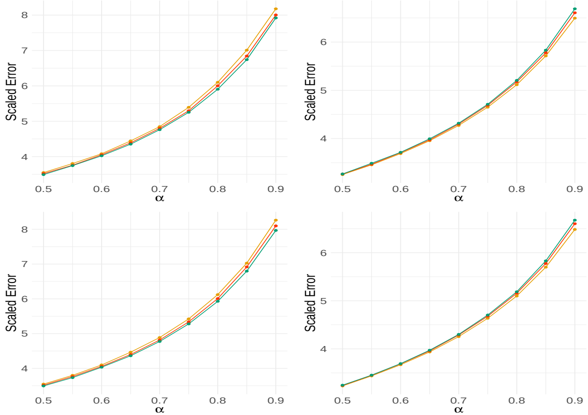


Figure: Top: SBM; Bottom: Graphon; Left: Adjacency; Right: Laplacian.

Outline

Change-point Problem Formulation

- Observe T graphs $G_1, \ldots, G_T$ on the same set of n vertices.
- $\mathbf{A}^{(t)}$ is the adjacency matrix of G_t.
- G_t comes from a lazy inhomogeneous Erdös-Rényi model with probability matrix, such that, there is an (unknown) time point $0 < \tau < T$ with

  $$\mathbb{E}(\mathbf{A}^{(t)}) = \mathbf{Q}_1 \text{ for } 1 \leq t \leq \tau \text{ and } \mathbb{E}(\mathbf{A}^{(t)}) = \mathbf{Q}_2 \text{ for } \tau{+}1 \leq t \leq T$$

  where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are unknown $n \times n$ probability matrices.
- $A_{ij}^{(\tau+1)}$ equals $A_{ij}^{(\tau)}$ with probability α, and is sampled independently (of $A^{(1:\tau)}$ and across i, j) from Bernoulli($\mathbf{Q}_2$) with probability $(1-\alpha)$.
- We want to estimate τ.

Parameters/Hyper-parameters

- T = Number of networks,
- κ_0 = (cushion) minimum gap between change point and the end points,
- $Sig = ||\mathbf{Q}_1 - \mathbf{Q}_2||$,
- n = Size of each network,
- d = Max expected degree of a node at a time ($d = n\rho$ where $\rho = \max_{i,j,t} \mathbf{P}_{i,j}^{(t)}$),
- (Algorithmic parameters)
  - κ = (buffer window length) target length of buffer from end points or between change points.

Single Change Point Detection

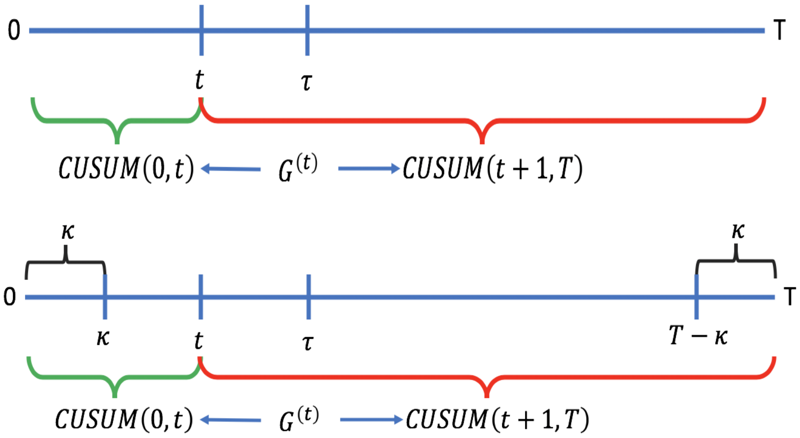


Figure: Pictorial Representation

The CUSUM Procedure for Single Change Point

1. Obtain average degree $\bar{D} = \frac{1}{nT} \sum_{i,j \in [n], s \in [T]} \mathbf{A}_{i,j}^{(s)}$.
2. Given κ, obtain the CUSUM statistic,

$$\mathbf{G}_{\xi}^{(t)} := \left(\frac{t}{T} \left(1 - \frac{t}{T} \right) \right)^{\xi} \left(\frac{1}{t} \sum_{s=1}^{t} \mathbf{A}^{(s)} - \frac{1}{T-t} \sum_{s=t+1}^{T} \mathbf{A}^{(s)} \right)$$

   for $\kappa \leq t \leq T - \kappa$.
3. Obtain $M := \max_t \left|\left|\mathbf{G}_{\xi}^{(t)}\right|\right|$, and potential change point estimate $\check{\tau} := \arg\max_t \left|\left|\mathbf{G}_{\xi}^{(t)}\right|\right|$.
4. If $M > C\sqrt{\frac{\bar{D}}{T}}$ for some specific constant C, then declare $\check{\tau}$ as a change-point estimate, $\hat{\tau}$.

Theoretical Results

The main theoretical result is the following nature -

Theorem (Chatterjee[2], Mukheree, Nath and B. (2022))

Let $\kappa_0 = \min\{\tau, T-\tau\}$. *For any* $c > 0$, *there exists a constant* $C > 0$ *such that for any* $\kappa \leq \kappa_0$, *if* $\kappa \min\{d_{\max}^{\mathbf{Q}_1}, d_{\max}^{\mathbf{Q}_2}\} > C(\log n)^3$, *then*

$$|\hat{\tau} - \tau| \leq C_4 \frac{T}{\left(1 - \frac{t}{T}\right)^{\xi} \|\mathbf{Q}_1 - \mathbf{Q}_2\|} \sqrt{\frac{\max\{d_{\max}^{\mathbf{Q}_1}, d_{\max}^{\mathbf{Q}_2}\}(1+C)\log(n)}{(1-\alpha)\kappa}}$$

with probability at least $1 - \frac{2T}{n^c}$.

Observations Regarding the Theorem

- The networks can only be $\log(n)$-sparse according to the current results. But, it might be possible to extend the current theoretical results.
- The result depends on the *lazyness* parameter, α and as expected as α increases the problem becomes more difficult as the number of independent layers decreases.

Simulation Results

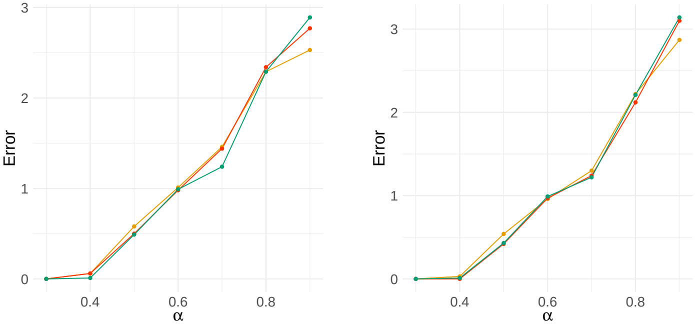


Figure: Average absolute error in change-point detection for graphon model over 100 experiments. Left panel: $\Delta_p = 0.05$. Right panel: $\Delta_p = 0.07$. Parameters: $\Delta_e = 100$ (yellow), 110 (red), 120 (green). We have used the CUSUM statistic with $\xi = 1/2$.

Outline

Community Detection Problem Formulation

- A sequence of networks $G^{(1)}, \ldots, G^{(T)}$ generated from the lazy IER process, with common mean matrix $P = ZBZ^\top$.
- $Z \in \{0,1\}^{n \times K}$ is the community membership matrix and $B \in [0,1]^{K \times K}$ is the matrix of between/within community edge-formation probabilities.
- **Goal:** Recover the community membership matrix Z.
- **Method:** Spectral clustering algorithm on (a) the aggregated adjacency matrix A gives $\hat{Z}_{\text{Adj}}$, and (b) aggregated Laplacian matrix $\mathcal{L}$ gives $\hat{Z}_{\text{Lap}}$.
- The mis-clustering error of a community estimator $\hat{Z}$ is defined to be -

$$\text{ME}(\hat{Z}, Z) := \inf_{\Pi \in \mathcal{P}_K} \frac{1}{n} \|\hat{Z} - Z\Pi\|_F^2,$$

where $\mathcal{P}_K$ is the set of all $K \times K$ permutation matrices.

Theoretical Results

Theorem (Chatterjee[2], Mukheree, Nath and B. (2022))

Let γ_{Adj} and γ_{Lap} denote the smallest non-zero singular value of $P = ZBZ^\top$ and $D_P^{-1/2} P D_P^{-1/2}$ respectively.
For any constant $c > 0$ there exists another constant $C = C(c) > 0$ such that if $Td_{\max} > C(\log(n))^3$, then there is a constant $C_1 = C_1(\alpha, c)$ such that for any $\delta \in (n^{-c}, 1/2)$,

$$ME(\hat{Z}_{Adj}, Z) \leq \frac{C_1(2+\epsilon)Kd_{\max}\log(n/\delta)}{\gamma_{Adj}^2 T}$$

with probability at least $1-\delta$. And,

$$ME(\hat{Z}_{Lap}, Z) \leq \frac{C_2(2+\epsilon)K\log(4n/\delta)}{\gamma_{Lap}^2 Td_{\min}}$$

with probability at least $1-\delta$.

The constants C_1 and C_2 are the same (up to absolute multiplicative constant) as their namesakes appearing in the main Theorem.

Simulation Results

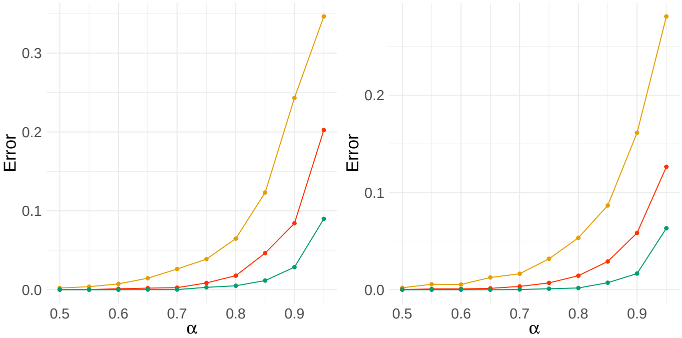


Figure: Average misclustering error in community detection for SBM over 100 experiments. Left panel: using the aggregated adjacency matrix. Right panel: using the aggregated Laplacian matrix. Parameters: $a = 7$, $b = 3$ (yellow); $a = 7.5$, $b = 2.5$ (red); $a = 8$, $b = 2$ (green).

Outline

Problem Formulation

- Observe T graphs $G_1, \ldots, G_T$ on the same set of n vertices.
- $\mathbf{A}^{(t)}$ is the adjacency matrix of G_t.
- $\mathcal{P} := \{\mathbf{P}^{(t)} : 1 \leq t \leq T\}$ is a stochastic process.
- G_t comes from an inhomogeneous Erdös-Rényi model with probability matrix: $\mathbb{E}\mathbf{A}^{(t)} = \mathbf{P}^{(t)}$.

Community Detection Problem Formulation

- $\mathbf{Z} \in \{0,1\}^{n\times K}$ is the community membership matrix and consider $\mathcal{B} := (\mathbf{B}^{(t)}_{K\times K}, t \in [T])$ is a stochastic process and $\mathbf{P}^{(t)} = \mathbf{Z}\mathbf{B}^{(t)}\mathbf{Z}^T$. $\mathcal{B}$ is the sequence of connection probability matrix between/within community.
- **Goal:** Recover the community membership matrix $\mathbf{Z}$.
- **Method:** Spectral clustering algorithm on the aggregated squared adjacency matrix $\mathbf{A}^{[2]} = \sum_{t=1}^{T}(\mathbf{A}^{(t)})^2 - \text{diag}(\sum_{t=1}^{T}(\mathbf{A}^{(t)})^2)$ gives $\hat{\mathbf{Z}}_0$.

Community Detection Algorithm

- Obtain $\mathbf{A}_0^{[2]} := \sum_{t=1}^{T} \left(\mathbf{A}^{(t)}\right)^2$ (sum of squares of the adjacency matrices) and zero out its diagonal to get $\langle \mathbf{A}_0^{[2]} \rangle$.
- Get $D_i^{[1]} := \max_{t=1}^{T} \sum_{j=1}^{n} A_{i,j}^{(t)}$ and $D_i^{[2]} := \sum_{j=1}^{n} \langle A_0^{[2]} \rangle_{i,j}$ for $i \in [n]$.
- Get the order statistics $D_{(1)}^{[1]} \leq \cdots \leq D_{(n)}^{[1]}$ and $D_{(1)}^{[2]} \leq \cdots \leq D_{(n)}^{[2]}$.
- Get $\bar{d}^2 := \frac{1}{nT} \sum_{i=1}^{n} D_i^{[2]}$. Get Γ_1 and Γ_2 as two thresholds.
- Get $\{i \in [n] : D_i^{[l]} \leq D_{(n+1-\Gamma_l)}^{[l]}$ for both $l = 1, 2\}$ and sort its entries in ascending order to have $1 \leq k_1 < \cdots < k_{n'} \leq n$.
- Get submatrix $\mathbf{A}^{[2]} \in \mathbb{R}^{n' \times n'}$ of $\langle \mathbf{A}_0^{[2]} \rangle$, where $A_{i,j}^{[2]} = \langle \mathbf{A}_0^{[2]} \rangle_{k_i, k_j}$.
- Perform Spectral Clustering on $\mathbf{A}^{[2]}$ and extend to obtain $\hat{\mathbf{Z}}_0$ as follows by including the excluded nodes in one community.
- $\hat{\mathbf{Z}}_0$ is the estimate of $\mathbf{Z}$.

Theoretical Results for Independent Layers

Theorem (B. and Chatterjee (2020))

For $a \in [K]$, let f_a denote the proportion of nodes having community label a, which are misclassified in Algorithm 1. For any $\epsilon > 0$, there are constants $C = C(\epsilon), c > 0$ such that if $\eta \in (0,1)$ is any number satisfying $\eta > c(Td(d \wedge 1))^{-1/3}$ and if

$$\lambda \left(\frac{n_{min}}{n}\right)^2 > \max\left\{\frac{7}{n}, \frac{C\sqrt{K}}{(Td(d \wedge 1))^{1/4}}\right\}, \quad (1)$$

then $\sum_{a \in [K]} f_a \le \frac{C^2 K \Psi}{\lambda^2 \left(\frac{n_{min}}{n}\right)^4}$ *with probability* $\ge 1 - \eta$, *where,*

$$\Psi := \begin{cases} (Td)^{-1/2} & \text{if } d > 1 \text{ and } n \ge (Td)^{1/4} \\ [n^2 + (n/6)\log(4/\eta)](Td)^{-1} & \text{if } d > 1 \text{ and } n < (Td)^{1/4} \\ (Td^2)^{-1/2} & \text{if } d \le 1 \text{ and } n \ge (Td)^{1/4} \\ [n^2 + (n/6)\log(4/\eta)](Td^2)^{-1} & \text{if } d \le 1 \text{ and } n < (Td)^{1/4} \end{cases}$$

Therefore, in the special case, when (i) K is a constant, (ii) the community sizes are balanced, i.e. $n_{max}/n_{min} = O(1)$, and (iii) the edge probabilities are of the same order, i.e. $\lambda \asymp 1$, then the proportion of misclassified nodes in $\hat{\mathbf{Z}}_0$ is arbitrarily small (resp. goes to zero) with high probability (resp. probability $1 - o(1)$) if $T(d \wedge d^2)$ is large enough (resp. $T(d \wedge d^2) \to \infty$).

Community Detection Problem Formulation in Dependent Case

- The smallest eigenvalue $\lambda_{K,t} := \lambda_K([\mathbf{B}^{(t)}]^2)$ with CDF $F_t(x) := \mathbb{P}(\lambda_{K,t} \leq x)$ for $x \geq 0$. Let $b_t := \mathbf{1}_{\{\lambda_{K,t}=0\}}$ and $F_t^+(x) := \frac{F_t(x)-F_t(0)}{1-F_t(0)}$, and $\tilde{\lambda}_{K,t} \sim F_t^+(x)$ for all t is an independent copy generated from the truncated distribution.
- The *maximal degree variable*, $\underline{d}_n(\epsilon)$ for any $\epsilon > 0$, is defined in the following way -

$$\underline{d}_n(\epsilon) := \sup\left\{x \in [0, n] : \mathbb{P}\left(\max_{t\in[T],\, a,b\in[K]} nB_{ab}^{(t)} \leq x\right) \leq \epsilon\right\}.$$

Community Detection Problem Formulation in Dependent Case

- **(Mixing condition)** Consider a decreasing function $\alpha_{\downarrow} : \mathbb{Z}_{+} \to [0, 1]$ to reflect the decay of correlation (at any rate) between two events of non-informative (smallest eigenvalue of $\mathbf{B}^{(t)}$ being zero) $\mathbf{B}^{(t)}$ matrices, like $\mathbf{B}^{(t_1)}$ and $\mathbf{B}^{(t_2)}$, where, $t_1, t_2 \in [T], t_1 \neq t_2$.

$$\left| \mathbb{P}\left(\cap_{i\in[2]} \{\lambda_{K,t_i} = 0\} \right) - \prod_{i\in[2]} \mathbb{P}(\lambda_{K,t_i} = 0) \right| \leq \alpha_{\downarrow}(|t_1 - t_2|) \qquad (2)$$

with $\alpha_{\downarrow}$ having the property

$$\alpha_{\downarrow}(s) \downarrow 0 \text{ as } s \uparrow \infty, \text{ and } \alpha_{\downarrow}(0) = 1.$$

This decay of correlation is necessary to have consistent recovery of communities.

Community Detection Problem Formulation in Dependent Case

- Consider a function $\psi_{\uparrow\downarrow} : \mathbb{N} \times \mathbb{R}_+ \to [0, 1]$ in terms of T and $\underline{d}_n(\epsilon)$, which captures the probability that network layers are non-informative, that is,

$$\max_{t \in [T]} \mathbb{P}\left(\{\lambda_{K,t} = 0\}\right) \leq \psi_{\uparrow\downarrow}(T, \underline{d}_n(\epsilon)). \tag{3}$$

  $\psi_{\uparrow\downarrow}(T,\underline{\text{d}})$ is a function which captures the behavior that on one hand $\psi_{\uparrow\downarrow}$ increases to 1 as T increases and $\underline{d}$ remains constant. But, on the other hand $\psi_{\uparrow\downarrow}$ decreases to 0 as $\underline{d}$ increases and the number of networks T stays the same, that is,

$$\lim_{T \uparrow \infty} \psi_{\uparrow\downarrow}(T, \underline{d}) = 1 \quad \text{and} \quad \psi_{\uparrow\downarrow}(T, \underline{d}) \downarrow 0 \text{ as } \underline{\text{d}} \uparrow \infty.$$

Community Detection Problem Formulation in Dependent Case

Based on the random variables $\lambda^+_{K,t}$ and $\underline{d}_n(\epsilon)$, and the functions $\alpha_\downarrow$, $\psi_{\uparrow\downarrow}$ and $\phi_\downarrow$, we place the following conditions on the stochastic process $\mathcal{B}$.
Assumption A: Let $\mathcal{B} = (\mathbf{B}^{(t)}, t \in [T])$ be a stochastic process with the following properties -

$$(a)\ \psi_{\uparrow\downarrow}(T, \underline{d}_n(\epsilon)) \leq 1 - \left[\frac{\sqrt{T}}{T} + \alpha_\downarrow(\sqrt{T})\right]^{1/2-\delta} \vee \frac{1}{(T\underline{d}_n(\epsilon)(\underline{d}_n(\epsilon) \wedge 1))^{\frac{1}{60}}}$$

$$(b)\ \max_{t\in[T]} F_t(0) \leq \psi_{\uparrow\downarrow}(T, \underline{d}_n(1/2)), \text{ and} \tag{4}$$

$$(c)\ \max_{t\in[T]} \mathbb{E}\phi_\downarrow(\lambda^+_{K,t}) \leq C_1 \text{ for a decreasing and convex function} \tag{5}$$

$$\phi_\downarrow : (0, \infty) \to (0, \infty)$$

for any $\epsilon > 0$ and for some constants $C_1 < \infty$ and $\delta < 1/2$.

Theoretical Results

Theorem (B. and Chatterjee (2020))

Let $(\mathbf{B}^{(t)}, t \in [T])$ *be any stochastic process satisfying Assumption A,* $(f_a, a \in [K])$ *be the mis-clustering error for each community. For any* $\epsilon > 0, \delta \in (0, 1/2)$,

$$\mathbb{P}\left(\sum_{a\in[K]} f_a > (T\underline{d}(\epsilon))^{-1/6}\right) \leq \epsilon + \frac{C_1}{\phi_{\downarrow}\left(\frac{2n^2}{n_{min}^2}\left[T\underline{d}_n(\epsilon)(\underline{d}_n(\epsilon)\wedge 1)\right]^{-1/15}\right)} \tag{6}$$

$$+\min\left\{4\left[\frac{\sqrt{T}}{T}+\alpha_{\downarrow}(\sqrt{T})\right]^{2\delta}, T\psi_{\uparrow\downarrow}(T, \underline{d}_n(\epsilon)(\underline{d}_n(\epsilon)\wedge 1))\right\}$$

$$+\frac{2C'+2nK}{n}\left(\left[T\underline{d}_n(\epsilon)(\underline{d}_n(\epsilon)\wedge 1)\right]^{-3/4}+\epsilon\right).$$

Therefore, in the special case, when (i) K *is a constant and (ii) the community sizes are balanced, i.e.* $n_{max}/n_{min} = O(1)$*, then the proportion of misclassified nodes in* $\hat{\mathbf{Z}}_0$ *is arbitrarily small (resp. goes to zero) with probability* $1 - o(1)$ *if* $T\underline{d}_n(\epsilon)(\underline{d}_n(\epsilon)\wedge 1)$ *is large enough (resp.* $T\underline{d}_n(\epsilon)(\underline{d}_n(\epsilon)\wedge 1) \to \infty$*) and* ϵ *is small enough (resp.* $\epsilon \to 0$*).*

Change-Point Problem Formulation

- Observe T graphs $G_1, \ldots, G_T$ on the same set of n vertices.
- $\mathbf{A}^{(t)}$ is the adjacency matrix of G_t.
- $\mathcal{P} := \{\mathbf{P}^{(t)} : 1 \leq t \leq T\}$ is a stochastic process.
- G_t comes from an inhomogeneous Erdös-Rényi model with probability matrix: $\mathbb{E}\mathbf{A}^{(t)} = \mathbf{P}^{(t)}$.
- Suppose that there is an (unknown) time point $0 < \tau < T$ such that

  $$\mathbb{E}(\mathbf{P}^{(t)}) = \mathbf{Q}_1 \text{ for } 1 \leq t \leq \tau \text{ and } \mathbb{E}(\mathbf{P}^{(t)}) = \mathbf{Q}_2 \text{ for } \tau{+}1 \leq t \leq T$$

  where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are unknown $n \times n$ probability matrices.
- We want to estimate τ.

Parameters/Hyper-parameters

- T = Number of networks,
- $Sig = ||\mathbf{Q}_1 - \mathbf{Q}_2||$,
- n = Size of each network,
- d = Max expected degree of a node at a time ($d = n\rho$ where $\rho = \max_{i,j,t} \mathbf{P}_{i,j}^{(t)}$),
- (Algorithmic parameters)
  - κ = (buffer window length) target length of buffer from end points or between change points.

Single Change Point Detection

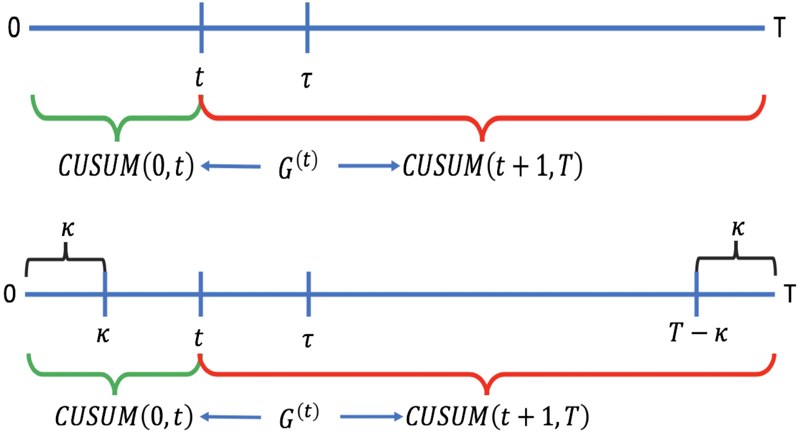


Figure: Pictorial Representation

The CUSUM Procedure for Single Change Point

1. Obtain average degree $\bar{D} = \frac{1}{nT}\sum_{i,j\in[n],s\in[T]} \mathbf{A}_{i,j}^{(s)}$.
2. Obtain submatrix $\tilde{\mathbf{A}}^{(s)}$ of $\mathbf{A}^{(s)}$ for all $s \in (0, T]$ by deleting rows (and corresponding columns) with high row sum.
3. Given κ, obtain the CUSUM statistic,

$$\mathbf{G}^{(t)} := \sqrt{\frac{t}{T}\left(1 - \frac{t}{T}\right)} \left(\frac{1}{t}\sum_{s=1}^{t} \tilde{\mathbf{A}}^{(s)} - \frac{1}{T-t}\sum_{s=t+1}^{T} \tilde{\mathbf{A}}^{(s)}\right)$$

   for $\kappa \leq t \leq T - \kappa$.
4. Obtain $M := \max_t \left\|\mathbf{G}^{(t)}\right\|$, and potential change point estimate $\check{\tau} := \arg\max_t \left\|\mathbf{G}^{(t)}\right\|$.
5. If $M > C\sqrt{\frac{\bar{D}}{T}}$ for some specific constant C, then declare $\check{\tau}$ as a change-point estimate.

Theoretical Results

The main theoretical result is the following nature -
Let $\mathcal{P}$ be stationary and ergodic process both before and after change point τ. There are constants $C, c > 0$ and vanishing sequence $\{\epsilon_T\}_T$ such that if $Sig \gg \sqrt{\frac{d}{T}}$, then

$$\mathbb{P}\left(|\hat{\tau} - \tau| = o(T)\right) \geq 1 - \exp\left(-\min\{C\log(n), cd\}\right) - \epsilon_T.$$

Some Observations

- The networks can be sparse. Maximum expected degree, d can be constant as well as going to infinity at an arbitrarily slow rate.
- The result follows from the use of Birkhoff's Ergodic Theorem. But, the result can be generalized by using mixing conditions.
- **Last note**, Carey Priebe and collaborators has done some excellent works in these directions albeit with a bit of different assumptions and models.

Outline

Bethe-Hessian Matrix

1. The Bethe Hessian matrix associated with an adjacency matrix $\mathbf{A}$ is defined as

$$\mathbf{H}_{\zeta} := (\zeta^2 - 1)\mathbf{I}_N + \mathbf{D} - \zeta\mathbf{A} \tag{7}$$

   where $\zeta > 1$ is a real scalar parameter, $\mathbf{D} := \mathrm{Diag}(\mathbf{A}\mathbf{1}_N)$.
2. Methods based on the spectrum of the Bethe Hessian operator have been considered empirically (e.g., Saade et al (2014), Dall'Amico et al (2019, 2020), and Le and Levina (2015)).
3. Le and Levina (2015) proved the consistency of the method based on the spectrum of the Bethe Hessian operator in semi-dense regimes, mean degree $d \gg \log n$.
4. However there is no theoretical result in the literature that guarantees the effectiveness of the Bethe Hessian operator in sparse regimes, $d = o(\log n)$.

Bethe-Hessian Matrix

1. Le and Levina (2015) showed that the number of informative eigenvalues of $\mathbf{H}_\zeta$ directly estimate K in the semi-dense regime ($\tilde{d} \gg \log(N)$) when ζ is set to be either r_m or r_a (degree based statistics). Both r_m and r_a are obtained based on heuristic arguments and are commonly used in the literature to estimate the radius of the bulk of the spectra.
2. (New Result) For a sparse regime when $1 \ll \tilde{d} \ll \log(N)$, let $\beta := \mathbf{B}_{\max}(1 - \lambda d_{\min} \check{N}_K)$, $\lambda := \lambda_K^{\downarrow}\left(\frac{N}{d}\mathbf{B}\right)$, $\check{N}_K := \sum_{i \in C_K} \psi_i^2 / \bar{d}_i$ and $d_{\min} := \min\left(\mathrm{Diag}(\bar{\mathbf{D}})\right)$. Then, for any $\delta \in (0, 3/2)$, $\mathbf{H}_\zeta$ has exactly K negative eigenvalues for all $\zeta \in \frac{1}{2}\left(-\beta \pm \sqrt{\beta^2 + 4 - 4d_{\min}}\right)$ with probability at least $1 - \exp[-(\zeta/\sqrt{d})^{3/2-\delta}]$.

Empirical K-estimation

- The theoretical result gives a theoretical interval but an empirical estimate within the interval is needed for practical applications.
- Following choice then works with high probability -
  1. Consider $\hat{d}_{\min} = \text{Median}\{d_i \quad | \quad d_i < \frac{1}{\sqrt{\log N}}\text{quantile}, i \in [N] \setminus \mathcal{R}\}$, where $\mathcal{R} = \{i \quad | \quad d_i \in \{0, 1\}\}$.
  2. Choose $\zeta = -\frac{\hat{\beta}}{2} = \sqrt{\hat{d}_{\min} - 1}$.
  3. Compute $\mathbf{H}_\zeta := (\zeta^2 - 1)\mathbf{I}_N + \mathbf{D} - \zeta\mathbf{A}$
  4. Perform eigenvalue decomposition of $\mathbf{H}_\zeta$ and let $\hat{\mathbf{K}}$ be the number of negative eigenvalues of $\mathbf{H}_\zeta$.
  5. $\hat{\mathbf{K}}$ is the estimate of $\mathbf{K}$

Consistency

Lemma (Hwang, Chatterjee, Xu, and B. (2022))

Let $\hat{d}_{\min} = \max\left\{d_i | d_i < \frac{c}{\sqrt{\log N}} \text{quantile}, i \in [N]\right\}$, *for some* $c \in (0, \sqrt{\log N})$. *Then, with probability at least* $1 - 2\delta$,

$$\hat{d}_{\min} \in \Big((1 - \varepsilon_\delta) d_{\min}, (1 + \varepsilon_\delta) d_{\min}\Big) \tag{8}$$

for any

$$\varepsilon_\delta \geq \sqrt{\frac{-2\log(c\delta/\sqrt{\log N})}{d_{\min}}} \tag{9}$$

Remark In Lemma , the quantity δ can be arbitrarily small depending on the concentration bound of the estimate $\hat{d}_{\min}$ to its population counterpart $d_{\min}$.

Theorem (Hwang, Chatterjee, Xu, and B. (2022))

Under the framework mentioned, with probability at least $1 - \delta$ *as defined in the Lemma,* $\hat{\mathbf{K}}$ *obtained from the Algorithm is the true* $\mathbf{K}$.

A Simulation Demonstration

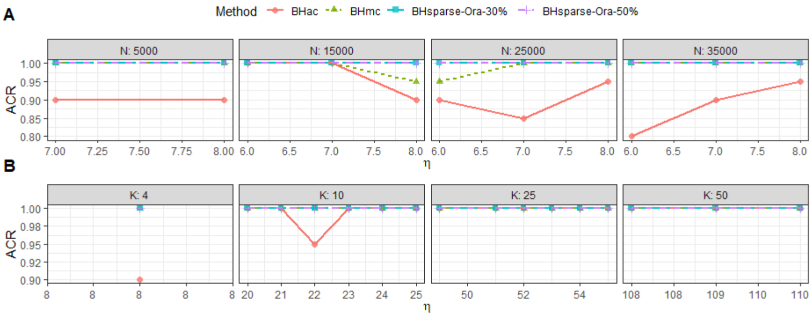


Figure: Row (A) shows Accuracy versus η using oracle intervals, with different values of N and $K = 3$. Row (B) shows ACR versus η as K varies with fixed $N = 25,000$. Both with fixed $\tilde{d} = 3\sqrt{\log(N)}$.

Outline

Five "Desirable" Properties of Complex Networks

- **Sparsity:** Average degree is bounded or grows slowly in terms of number of vertices.
- **Small-world property:** The diameter of the network is bounded or grows slowly in terms of number of vertices.
- **Power-law degree distribution:** The degree distribution of the networks follow scale-free power-law models.
- **Transitivity:** The network structure has transitive property, that is, the average number of triangles grows in terms of number of vertices.
- **Clusterability or presence of community or low-rank structure:** Possibility of the presence of a community structure in the networks.

Network Models with all "Five" Properties

- Graphon or inhomogeneous Erdös-Rényi random graph models can usually have at max four of these properties - it is difficult to attain both sparse, transitive and low-rank networks from graphon models as they have conditionally independent edges.
- Exponential random graph models have the possibility of generating networks with all five properties but are difficult to analyze.
- Other network models, such as, edge-exchangeable network models (Caron and Fox, 2015; Crane and Dempsey, 2015) have been proposed which also have the possibility of having all five properties.

Outline

Transitive Inhomogeneous Erdös-Rényi (TIER) model

Transitive inhomogeneous Erdös-Rényi (TIER) model of k closures, TIER(k), can be described in terms of two parameters:

- connectivity probability matrix $\mathbf{P}_{n \times n}$,
- triangle closure parameter δ_n,
- number of closures, k.

The TIER graph generating mechanism is given by -

1. Start with $\ell = 0$, and $\mathbb{P}\left(\tilde{A}_{ij} = 1\right) = \mathbf{P}_{ij}$, (inhomogeneous Erdös-Rényi generation step),
2. If $\tilde{A}_{ij} = 0$, then $\tilde{n}_{ij} := \sum_{k \neq i,j} \tilde{A}_{ik}\tilde{A}_{jk}$ (detection of 'V' structures)
3. $\mathbb{P}(A_{ij} = 1) = \min(1, \delta_n)\mathbf{1}(\tilde{n}_{ij} > 0)$, and $\ell = \ell + 1$ (triangle or transitive closure)
4. If $\ell < k$, go back to step 2 with $\tilde{\mathbf{A}} = \mathbf{A}$. Otherwise, $\mathbf{A} = (A_{ij})_{i,j=1}^{n}$ is the generated adjacency matrix.

TIER model Comments

- TIER(k) model extends the graphon or IER model set-up to allow for dependent edge formation using transitive closure. Networks generated from the TIER model are not conditionally independent.
- Transitive closures are common for directed graphs and have also been applied to extend Chung-Lu models (transitive Chung-Lu model, Pfeiffer et.al. (2012)).
- TIER(k) models can potentially generate networks with all five properties depending on the generating graphon.

Another Related Work

- Yuan and Qu (2021) proposed a conditional stochastic block model with correlation structure on the adjacency matrix.
- The work was based on Bahadur's approximation of multivariate Bernoulli distributions and has close ties to the transitive closures of TIER models.

Properties of an Example of TIER model

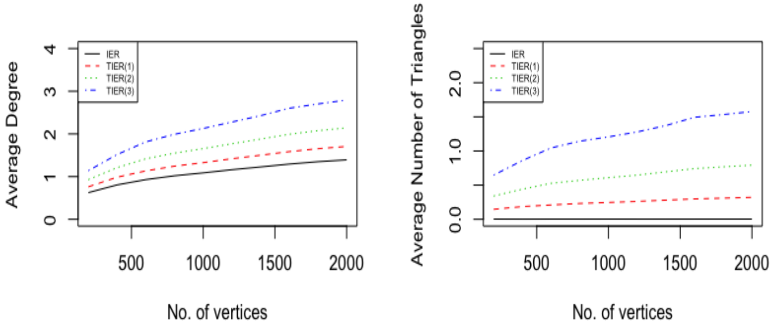


Figure: TIER models generated using stochastic block model kernels.

Properties of an Example of TIER model

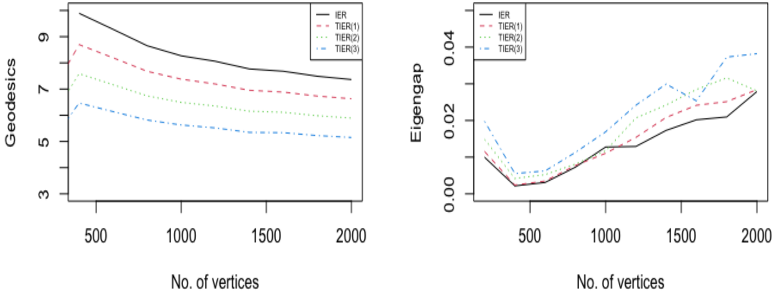


Figure: TIER models generated using stochastic block model kernels.

TIER model Comments

- Networks generated from TIER(k) models can be studied in several network inference problems -
  1. Subgraph count distributions,
  2. Community detection,
  3. Change point detection,
  4. Link prediction.
- However, we shall focus on a specific change point detection problem.
- We are working on Problems 1 and 3. With Avanti Athreya, Vince Lyzinski, and Jesus Arroyo, we are working on Problem 2.

Outline

Parameters/Hyper-parameters

- T = Number of networks,
- n = Size of each network,
- d = Max expected degree of a node at a time ($d = n\rho$ where $\rho = \max_{i,j,t} \mathbf{P}_{i,j}^{(t)}$),
- δ_n = transitive closure probability,
- (Algorithmic parameter)
    - κ = (buffer window length) target length of buffer from end points or between change points.
    - γ = (group size) the size of non-overlapping local groups.
    - H is the subgraph of interest, most commonly, triangles.

The Procedure for Local Change Point Detection

1. Divide the set of nodes randomly into g non-overlapping groups, $\{V_1, V_2, \dots, V_g\}$, such that, $g\gamma = n$.
2. For each node $i \in \{1, \dots, n\}$, compute, $N_i^{(H,t)}$, the count of subgraphs, H, containing node i.
3. For each group $h \in \{1, \dots, g\}$, for each $t \in (\kappa, T - \kappa)$, compute,

$$S_h^{(H,t)} = \frac{1}{|V_h|} \sum_{i \in V_h} N_i^{(H,t)}$$

4. Obtain the CUSUM statistic, for each group $h \in \{1, \dots, g\}$ and $t \in (\kappa, T - \kappa)$,

$$G_h^{(H,t)} := \sqrt{\frac{t}{T}\left(1 - \frac{t}{T}\right)} \left(\frac{1}{t} \sum_{s=1}^{t} S_h^{(H,s)} - \frac{1}{T-t} \sum_{s=t+1}^{T} \tilde{S}_h^{(H,s)} \right).$$

The Procedure for Local Change Point Detection II

5. Obtain for each group $h \in \{1, \ldots, g\}$, $M_h := \max_{t \in (\kappa, T-\kappa)} \left| G_h^{(H,t)} \right|$, and potential change point estimate $\check{\tau}_h := \arg\max_{t \in (\kappa, T-\kappa)} \left| G_h^{(H,t)} \right|$.
   For each group $h \in \{1, \ldots, g\}$, if $M_h > A_h$ for some threshold A_h estimated using normal approximation, then declare $\check{\tau}_h$ as a change-point estimate for group h.

Comments on the Local Algorithm

- The number of groups, g, can range between $O(1)$ to $O(n)$ to detect change points in different multi-scale localities.
- The change-point locations can be different for different groups or channels.
- The procedure can be used for multiple change point detection using binary segmentation procedures.
- The procedure can be extended for online change point detection too.

Theoretical Results for Local Change point Detection

- We have derived theoretical results for some special cases only for now.
- For the subgraph, H, as triangle, the results depend on the asymptotic distribution of the triangle counts in TIER model networks. The asymptotic distribution can be derived in several different cases -
  - **(Sparse Case)** For $d\delta_n \to 0$ and $d^2\delta_n \to \infty$, normalised local triangle count statistic has asymptotic standard normal distribution. The proof follows using Stein's method.
  - **(dense case)** For $d\delta_n \to \infty$ and and $d^2\delta_n \to \infty$, normalised local triangle count statistic also has asymptotic standard normal distribution. But, the proof uses U-statistic method.

Observations Regarding the Theoretical Results

- The threshold for change point detection, A_h, can either be derived using *Brownian bridge* approximation or *simulated* distributions. In current case, we use simulated distributions.
- Depending on the *sparsity* of the networks, the size of the groups can be decided. For very sparse networks, the group size has to be large, but for dense networks, it can be arbitrarily small.
- Instead of triangles, counts for other subgraphs can also be used, but the theoretical results for general subgraph counts are tedious and still under investigation.

Simulation Results

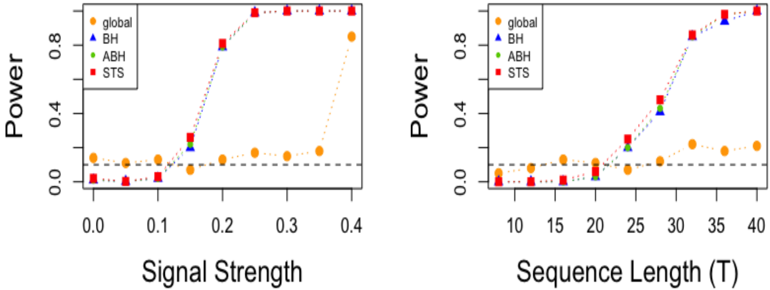


Figure: Power of the local change point detection methods.

Results for US Senate Networks

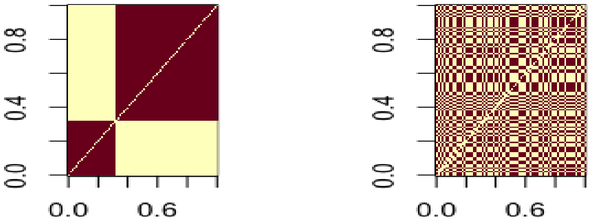


Figure: Change in US Senate Network at change-point of Jan 10, 1994. The local change point detects the change point for groups 2, 4, 7, and 10 for 10 consecutive groups each of size 10. These 4 groups have the highest change proportions.

Outline

Future directions

- Address the extensions in TIER(k) models.
- Extended the dependence set-up of dependent adjacency matrices.
- Generalization to the case where the vertex set may evolve with time and/or there are more general dependence structures among the edges.
- How to incorporate available information (covariates) about the networks to improve the methods?

An Advertisement

Two tenure-track openings at Department of Statistics

Figure: Oregon State University

References

1. Bhattacharyya, S. and Chatterjee, S., 2020. General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers. arXiv preprint arXiv:2004.03480. Under Revision in AOS.
2. Bhattacharyya, S., Chatterjee, S. and Mukherjee, S.S., 2020. Consistent detection and optimal localization of all detectable change points in piecewise stationary arbitrarily sparse network-sequences. arXiv preprint arXiv:2009.02112.
3. Hwang, N., Xu, J., Chatterjee, S. and Bhattacharyya, S., 2020. Estimation of Number of Communities in Assortative Sparse Networks. Under Submission in JASA.
4. Hwang, N., Xu, J., Chatterjee, S. and Bhattacharyya, S., 2021. The Bethe Hessian and Information Theoretic Approaches for Online Change-Point Detection in Network Data. Sankhya, 2021.

Thank You!