Changepoint Estimation in Sparse Dynamic Stochastic Block Models under Near-Optimal Signal Strength

Shirshendu Chatterjee[1], Soumendu Sundar Mukherjee[2] & **Tamojit Sadhukhan**[2]
[1]City University of New York, New York [2]Indian Statistical Institute, Kolkata

Introduction

- Consider T independent networks on n nodes with adjacency matrices $A^{(1)}, \ldots, A^{(T)}$.
- There are K communities and $\left(B^{(t)}_{K\times K}\right)_{t\in[T]}$ are community-wise block-connectivity matrices.
- The community-label of node $i \in [n]$ is $z_i \in [K]$. For $t \in [T]$ and $i, j \in [n]$ with $i < j$,

$$P^{(t)}_{ij} := \mathbb{P}\big(A^{(t)}_{ij} = 1\big) = B^{(t)}_{z_i z_j}.$$

- Let $Z_{n\times K}$ be given by $Z_{ik} = \mathbf{1}_{\{z_i = k\}}$, so that for $t \in [T]$, $P^{(t)} = ZB^{(t)}Z^T$.
- The block-connectivity matrices $B^{(t)}, t \in [T]$ change at some $\tau \in [T]$:

$$B^{(t)} = \begin{cases} B_0 & \text{if } 1 \le t \le \tau, \\ B_1 & \text{if } \tau + 1 \le t \le T. \end{cases}$$

- Thus at the changepoint $\tau \in [T]$ -

(a) some of the existing communities can split into multiple communities,
(b) the connection probabilities between some pairs of communities can change,
(c) some of the existing communities can merge into bigger communities,
(d) some of the communities can split and some of the communities can merge simultaneously.

- Our goal is to estimate the location of the changepoint τ consistently.

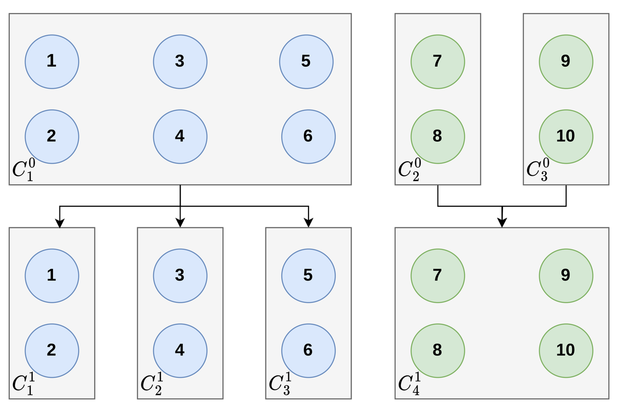

Figure 1: An example of simultaneous merging and splitting - there are 10 nodes with three communities C^0_1, C^0_2, C^0_3 before τ and four communities $C^1_1, C^1_2, C^1_3, C^1_4$ after τ.

$$B_0 = \begin{pmatrix} p & p & p & q & q \\ p & p & p & q & q \\ p & p & p & q & q \\ q & q & q & p & q \\ q & q & q & q & p \end{pmatrix}; \quad B_1 = \begin{pmatrix} p & q & q & q & q \\ q & p & q & q & q \\ q & q & p & q & q \\ q & q & q & p & p \\ q & q & q & p & p \end{pmatrix};$$

$$Z = \begin{pmatrix} \mathbf{1}_m e_1^\top \\ \mathbf{1}_m e_2^\top \\ \mathbf{1}_m e_3^\top \\ \mathbf{1}_m e_4^\top \\ \mathbf{1}_m e_5^\top \end{pmatrix}.$$

Our Contributions

- Complexity of the problem depends on a) the sparsity of the networks $d := n\,(\|B_0\|_\infty \vee \|B_1\|_\infty)$, b) the signal strength $\|B_0 - B_1\|_F$, and c) the cushion $\kappa = \min\{\tau, T - \tau\}$.
- We obtain a lower bound result that shows no algorithm can estimate τ consistently if

$$\|B_0 - B_1\|_F = O\left(\sqrt{\left(\frac{K}{n}\right)^3 \frac{d}{\kappa}}\right).$$

- We obtain a matching (when K does not grow with n) upper bound by devising a polynomial-time changepoint estimation algorithm. For $\Lambda \in [T]$ and $L = \lceil 3T/\Lambda \rceil - 2$, our algorithm can detect τ consistently and give a high probability confidence interval of length Λ if

$$Td \gg 1 \quad \text{and} \quad \|B_0 - B_1\|_F \gg \sqrt{K \log(KL)}\sqrt{\left(\frac{K}{n}\right)^3 \frac{d}{\Lambda \wedge \kappa}}.$$

- Existing methods require either higher signal strength (Bhattacharyya et al. (2020)), or the networks to be dense (Bhattacharjee et al. (2020)), or both (Zhao et al. (2019), Wang et al. (2021)).

Changepoint Estimation Method

Input: $\left(A^{(t)}\right)_{t\in[T]}, \delta \in [0,1], C_s > 0, \Lambda_{\min}$.

0. Obtain $\widehat{Z}$ using Algorithm 1 (Spectral Clustering of the Sum of Adjacency Matrices) of Bhattacharyya and Chatterjee (2020).

1. For all $m \in [K]$, set $\widehat{C}_m = \{k : \widehat{Z}_{km} = 1\}$ and $\widehat{n}_m = |\widehat{C}_m|$.

2. For all $t \in [T]$ and $k, l \in [K]$, obtain

$$\widehat{B}^{(t)}_{kl} = \frac{1}{\widehat{n}^2_{kl}} \sum_{\substack{i\in\widehat{C}_k, j\in\widehat{C}_l \\ i<j}} A^{(t)}_{ij}, \text{ where } \widehat{n}^2_{kl} = \begin{cases} \binom{\widehat{n}_k}{2} & \text{if } k = l, \\ \widehat{n}_k \widehat{n}_l & \text{if } k \ne l. \end{cases}$$

3. Set $\widehat{d} = \frac{1}{nT} \sum_{t\in[T]} \sum_{i,j\in[n]} A^{(t)}_{ij}$.

4. For each $\bar{\Lambda} = T, \ldots, \Lambda_{\min}$, repeat steps 5, 6, 7.

5. For $l = 1, \ldots, L(\bar{\Lambda}) := \lceil 3T/\bar{\Lambda} \rceil - 2$ and for $u \in \widetilde{S}(\bar{\Lambda}) := \left[\lfloor \bar{\Lambda}/3 \rfloor + 1, \ldots, \bar{\Lambda} - \lfloor \bar{\Lambda}/3 \rfloor\right]$, set

$$T_l = T_l(\bar{\Lambda}) := (l-1)\lfloor \bar{\Lambda}/3 \rfloor, \quad \mathcal{I}_l(\bar{\Lambda}) = (T_l(\bar{\Lambda}), T_l(\bar{\Lambda}) + \bar{\Lambda}]$$

and evaluate the CUSUM statistics for the window $\mathcal{I}_l(\bar{\Lambda})$ -

$$\widehat{\boldsymbol{G}}^{(T_l+u)}_l(\bar{\Lambda}) := \left(\frac{u}{\bar{\Lambda}}\left(1 - \frac{u}{\bar{\Lambda}}\right)\right)^{\delta} \left(\frac{1}{u}\sum_{v=1}^{u} \widehat{B}^{(T_l+v)} - \frac{1}{\bar{\Lambda} - u}\sum_{v=u+1}^{\bar{\Lambda}} \widehat{B}^{(T_l+v)}\right). \qquad (1)$$

6. For all $l = 1, \ldots, L(\bar{\Lambda}) := \lceil 3T/\bar{\Lambda} \rceil - 2$, obtain a candidate estimate in the window $\mathcal{I}_l(\bar{\Lambda})$ -

$$\tau_l = T_l(\bar{\Lambda}) + \underset{u\in\widetilde{S}(\bar{\Lambda})}{\operatorname{argmax}} \left\|\widehat{\boldsymbol{G}}^{(T_l+u)}_l(\bar{\Lambda})\right\|_F.$$

7. If the maximum of norm of CUSUM statistics evaluated at these candidates exceeds a threshold

$$\max_{l\in[L(\bar{\Lambda})]} \left\|\widehat{\boldsymbol{G}}^{(\tau_l)}_l(\bar{\Lambda})\right\|_F > C_s\sqrt{\left(\frac{K}{n}\right)^3 \frac{\widehat{d}}{\bar{\Lambda}}},$$

set the maximizing candidate as the changepoint estimate corresponding to $\bar{\Lambda}$ -

$$l(\bar{\Lambda}) = \underset{l\in[L(\bar{\Lambda})]}{\operatorname{argmax}} \left\|\widehat{\boldsymbol{G}}^{(\tau_l)}_l(\bar{\Lambda})\right\|_F \quad \text{and} \quad \widehat{\tau}(\bar{\Lambda}) = \tau_{l(\bar{\Lambda})}.$$

8. Let $\bar{\Lambda}_0$ be the minimum $\bar{\Lambda}$ for which $\widehat{\tau}(\bar{\Lambda})$ is obtained. Declare $\widehat{\tau} := \widehat{\tau}(\bar{\Lambda}_0)$ with

$$\left[(l(\bar{\Lambda}_0) - 1)\left\lfloor\frac{\bar{\Lambda}_0}{3}\right\rfloor + 1, (l(\bar{\Lambda}_0) - 1)\left\lfloor\frac{\bar{\Lambda}_0}{3}\right\rfloor + \bar{\Lambda}_0\right]$$

as a confidence interval for τ.

Simulation Experiments

- We compare the performance of our method under setups (a), (b), (c), (d) across varying sparsity.

Competing Methods	$\|B_0 - B_1\|_F \left(\left(\frac{K}{n}\right)^3 \frac{d}{\kappa}\right)^{-1/2}$	d
Bhattacharjee et al. (2020)	$\sqrt{KT/\kappa}$	$n^{3/5}$
Wang et al. (2021)	$\sqrt{n}\log^{1+\epsilon}(T)$	$\log n$
Zhao et al. (2019)	$\left(n^3 T/K^2\right)^{1/4}$ polylog(n)	cn
Bhattacharyya et al. (2020)	$\sqrt{n/K}$	T^{-1}
Our Method	$\sqrt{K \log K}$	T^{-1}

Table 1: Signal Strength and Sparsity Required for Competing Methods

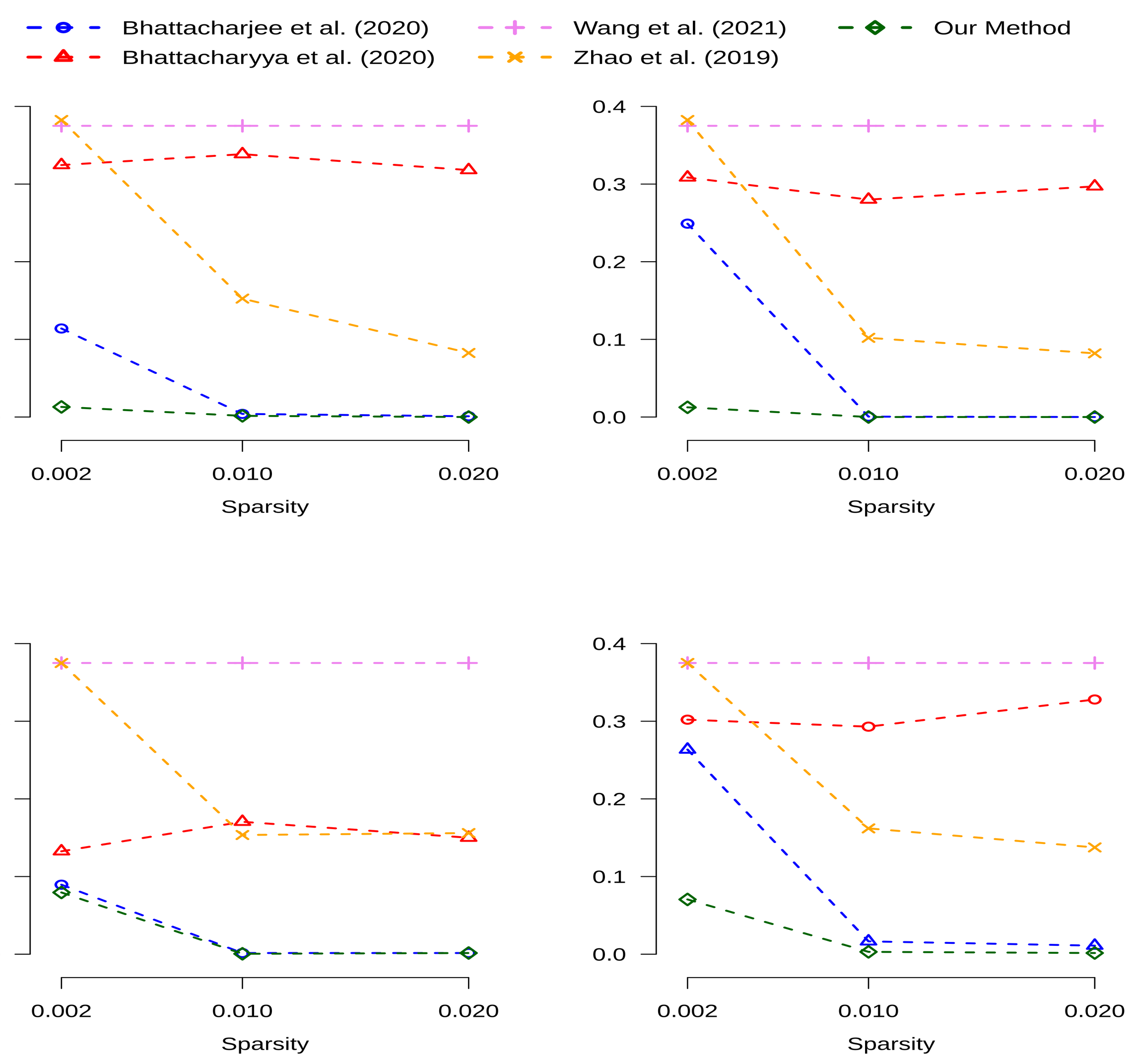

Figure 2: We take $\delta = 0.5$, $C_s = 6$ and $\Lambda_{\min} = 100$ and perform 10 Monte-Carlo (MC) runs for each setup. The mean (across 10 MC runs) relative errors in each setup (top row: (a),(b), bottom row: (c),(d)) for every method are shown.

A Real-World Example of Community Splitting

- US Senate roll call data (1979 - 2023) (Lewis et al. (2024)) illustrates community splitting.
- Representing Senate seats as nodes, 49 consecutive roll calls are used to construct a single network, resulting in $n = 100$ and $T = 199$. In each network, there is an edge between two seats if their stand on proposed bill is same in at least $\sim 70\%$ of the 49 roll calls.

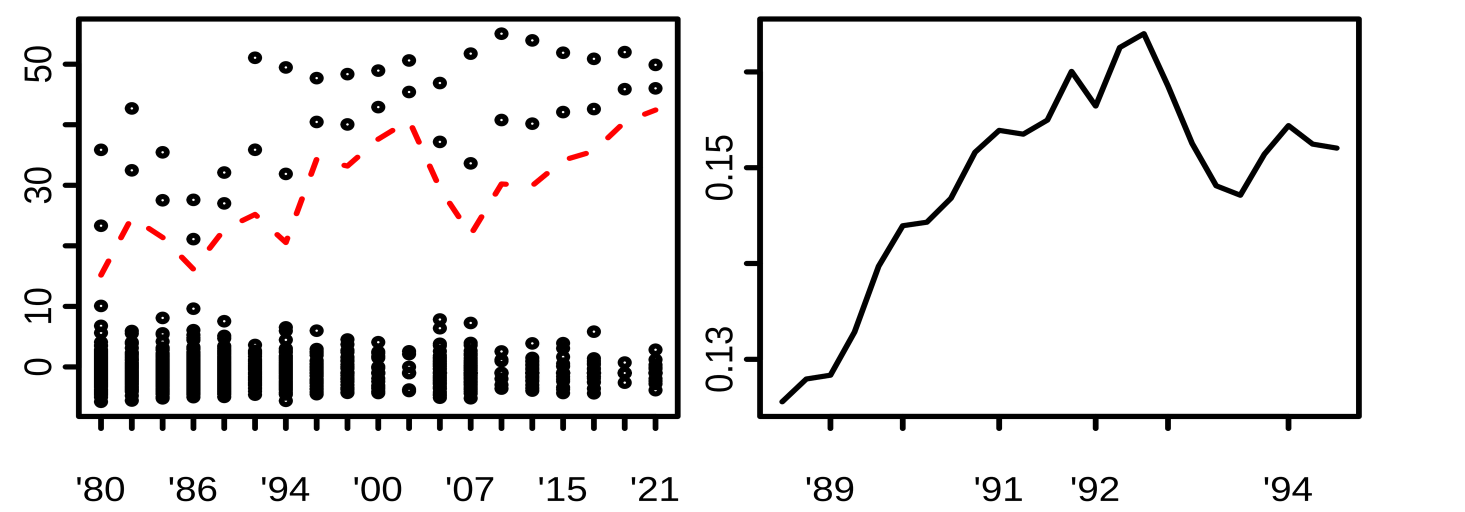

Figure 3: Left: Eigenvalue distributions of the roll call networks across time - two eigenvalues stick outside of the bulk spectrum, hinting at the presence of two communities. The dashed line shows the ratio of the squared second largest eigenvalue and the largest eigenvalue of these networks, a measure of the strength of the community structure - an overall increasing trend in this measure aligns with the increase in party polarization in US politics (Moody and Mucha (2013)). Right: Observed CUSUM statistics (1) (with $l = 2$ and $\Lambda = 70$) for the period $1988 - 1994$. The peak (estimated changepoint) corresponds to 49 consecutive roll calls from September 1992 to February 1993 which is near the November 1994 election when the Republican Party regained control of the House of Representatives for the first time since 1956. A sharp increase in party polarisation had been reported around this period in Moody and Mucha (2013).

- Temporal variations in the eigenvalue distributions suggest the presence of multiple changepoints.
- So, we combine our method for single changepoint estimation with wild binary segmentation (Fryzlewicz et al. (2014)) to detect several potentially important changepoints in these networks.

Threshold (C_s)	Detected Changepoints
60	62
50	40, 62, 125
40	40, 62, 71, 116, 125, 133, 158

Table 2: Wild Binary Segmentation

Detected Changepoints	Time-Period
40	1986/09 - 1986/10
62	1992/09 - 1993/02
71	1994/07 - 1994/08
116	2004/05 - 2004/09
125	2006/07 - 2007/01
133	2008/05 - 2009/01
158	2014/12 - 2014/12

Table 3: Time-Period Corresp. to Detected Changepoints

Future Directions

- Our method is not minimax-optimal when K grows with n.
- Our method needs to be improvised when there are multiple changepoints, or individuals switch communities at changepoints, or there are correlations among the network layers.
- Further, it will be of interest to extend our methodology to the setting of online changepoints.

Acknowledgements

The authors thank the anonymous reviewers for extensive and helpful comments. The research of the first author was supported by NSF grant DMS-2154564. The second author was partially supported by the INSPIRE research grant DST/INSPIRE/04/2018/002193 from the Dept. of Science and Technology, Govt. of India, and a Start-Up Grant from Indian Statistical Institute. The third author is partially supported by the Breakout Graduate Fellowship from International Mathematical Union (IMU).