Mu *et al. Applied Network Science* (2023) 8:5
https://doi.org/10.1007/s41109-022-00528-1

Applied Network Science

RESEARCH Open Access

Dynamic network sampling for community detection

Cong Mu[1*], Youngser Park[2] and Carey E. Priebe[1]

*Correspondence:
cmu2@jhu.edu

[1] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA
[2] Center for Imaging Science, Johns Hopkins University, Baltimore, USA

Abstract

We propose a dynamic network sampling scheme to optimize block recovery for stochastic blockmodel in the case where it is prohibitively expensive to observe the entire graph. Theoretically, we provide justification of our proposed Chernoff-optimal dynamic sampling scheme via the Chernoff information. Practically, we evaluate the performance, in terms of block recovery, of our method on several real datasets from different domains. Both theoretically and practically results suggest that our method can identify vertices that have the most impact on block structure so that one can only check whether there are edges between them to save significant resources but still recover the block structure.

Keywords: Dynamic network sampling, Stochastic blockmodel, Community detection, Chernoff information

Introduction

In network inference applications, it is important to detect community structure, i.e., cluster vertices into potential blocks. However, it can be prohibitively expensive to observe the entire graph in many cases, especially for large graphs. For example, in a network where vertices represent landline phones and edges represent whether there is a call between two landline phones. Based on the size of the network, in terms of the number of vertices, it can be extremely expensive to check whether there is a call for every landline phone pairs. Therefore, if one can utilize the information carried by a partially oberverd graph, that is only a small number of landline phone pairs are verified, to identify the landline phones that may play a more important role in formulating communities. Then given limited resources, one can choose to only check whether there are calls between those landline phone pairs to achieve the goal of detecting potential block structure. Thus it becomes essential to identify vertices that have the most impact on block structure and only check whether there are edges between them to save significant resources but still recover the block structure.

Many classical methods only consider the adjacency or Laplacian matrices for community detection (Fortunato and Hric 2016). By contrast, vertex covariates can also be taken into consideration for the inference. These covariate-aware methods rely on either variational methods (Choi et al. 2012; Roy et al. 2019; Sweet 2015) or spectral

approaches (Binkiewicz et al. 2017; Huang and Feng 2018; Mele et al. 2022; Mu et al. 2022). However, none of them focus on the problem of clustering vertices for partially observed graphs. To address this issue, existing methods propose different types of random and adaptive sampling strategies to minimize the information loss from the data reduction (Yun and Proutiere 2014; Purohit et al. 2017).

We propose a dynamic network sampling scheme to optimize block recovery for stochastic blockmodel (SBM) when we only have limited resources to check whether there are edges between certain selected vertices. The innovation of our approach is the application of Chernoff information. To our knowledge, this is the first time that it has been applied to network sampling problems. Motivated by the Chernoff analysis, we not only propose a dynamic network sampling scheme to optimize block recovery, but also provide the framework and justification for using Chernoff information in subsequent inference for graphs.

The structure of this article is summarized as follows. Section 2 reviews relevant models for random graphs and the basic idea of spectral methods. Section 3 introduces the notion of Chernoff analysis for analytically measuring the performance of block recovery. Section 4 includes our dynamic network sampling scheme and theoretical results. Section 5 provides simulations and real data experiments to measure the algorithms' performance in terms of actual block recovery results. Section 6 discusses the findings and presents some open questions for further investigation. Appendix provides technical details for our theoretical results.

Models and spectral methods

In this work, we are interested in the inference task of block recovery (community detection). To model the block structure in edge-independent random graphs, we focus on the SBM and the generalized random dot product graph (GRDPG).

Definition 1 (*Generalized Random Dot Product Graph* Rubin-Delanchy et al. 2022) Let $\mathbf{I}_{d_+d_-} = \mathbf{I}_{d_+} \bigoplus \left(-\mathbf{I}_{d_-}\right)$ with $d_+ \geq 1$ and $d_- \geq 0$. Let F be a d-dimensional inner product distirbution with $d = d_+ + d_-$ on $\mathcal{X} \subset \mathbb{R}^d$ satisfying $\mathbf{x}^\top \mathbf{I}_{d_+d_-} \mathbf{y} \in [0,1]$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Let $\mathbf{A}$ be an adjacency matrix and $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_n]^\top \in \mathbb{R}^{n \times d}$ where $\mathbf{X}_i \sim F$, i.i.d. for all $i \in \{1, \cdots, n\}$. Then we say $(\mathbf{A}, \mathbf{X}) \sim \text{GRDPG}(n, F, d_+, d_-)$ if for any $i, j \in \{1, \cdots, n\}$

$$\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{P}_{ij}) \qquad \text{where} \qquad \mathbf{P}_{ij} = \mathbf{X}_i^\top \mathbf{I}_{d_+d_-} \mathbf{X}_j. \tag{1}$$

Definition 2 (*K-block Stochastic Blockmodel Graph* Holland et al. 1983) The K-block stochastic blockmodel (SBM) graph is an edge-independent random graph with each vertex belonging to one of K blocks. It can be parametrized by a block connectivity probability matrix $\mathbf{B} \in (0,1)^{K \times K}$ and a vector of block assignment probabilities $\boldsymbol{\pi} \in (0,1)^K$ summing to unity. Let $\mathbf{A}$ be an adjacency matrix and $\boldsymbol{\tau}$ be a vector of block assignments with $\tau_i = k$ if vertex i is in block k (occuring with probability π_k). We say $(\mathbf{A}, \boldsymbol{\tau}) \sim \text{SBM}(n, \mathbf{B}, \boldsymbol{\pi})$ if for any $i, j \in \{1, \cdots, n\}$

$$\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{P}_{ij}) \qquad \text{where} \qquad \mathbf{P}_{ij} = \mathbf{B}_{\tau_i \tau_j}. \tag{2}$$

Remark 1

The SBM is a special case of the GRDPG model. Let $(\mathbf{A}, \boldsymbol{\tau}) \sim \text{SBM}(n, \mathbf{B}, \boldsymbol{\pi})$ *as in Definition 2 where* $\mathbf{B} \in (0,1)^{K\times K}$ *with* d_+ *strictly positive eigenvalues and* d_- *strictly negative eigenvalues. To represent this SBM in the GRDPG model, we can choose* $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_K \in \mathbb{R}^d$ *where* $d = d_+ + d_-$ *such that* $\boldsymbol{\nu}_k^\top \mathbf{I}_{d_+ d_-} \boldsymbol{\nu}_\ell = \mathbf{B}_{k\ell}$ *for all* $k, \ell \in \{1, \cdots, K\}$. *For example, we can take* $\boldsymbol{\nu} = \mathbf{U}_B |\mathbf{S}_B|^{1/2}$ *where* $\mathbf{B} = \mathbf{U}_B \mathbf{S}_B \mathbf{U}_B^\top$ *is the spectral decomposition of* $\mathbf{B}$ *after reordering. Then we have the latent position of vertex i as* $\mathbf{X}_i = \boldsymbol{\nu}_k$ *if* $\tau_i = k$.

The parameters of the models can be estimated via spectral methods (Von Luxburg 2007), which have been widely used in random graph models for community detection (Lyzinski et al. 2014, 2016; McSherry 2001; Rohe et al. 2011). Two particular spectral embedding methods, adjacency spectral embedding (ASE) and Laplacian spectral embedding (LSE), are popular since they enjoy nice propertices including consistency (Sussman et al. 2012) and asymptotic normality (Athreya et al. 2016; Tang and Priebe 2018).

Definition 3 (*Adjacency Spectral Embedding*) Let $\mathbf{A} \in \{0,1\}^{n\times n}$ be an adjacency matrix with eigendecomposition $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ where $|\lambda_1| \geq \cdots \geq |\lambda_n|$ are the magnitude-ordered eigenvalues and $\mathbf{u}_1, \cdots, \mathbf{u}_n$ are the corresponding orthonormal eigenvectors. Given the embedding dimension $d < n$, the adjacency spectral embedding (ASE) of $\mathbf{A}$ into $\mathbb{R}^d$ is the $n \times d$ matrix $\widehat{\mathbf{X}} = \mathbf{U}_A |\mathbf{S}_A|^{1/2}$ where $\mathbf{S}_A = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and $\mathbf{U}_A = [\mathbf{u}_1 | \cdots | \mathbf{u}_d]$.

Remark 2

There are different methods for choosing the embedding dimension (Hastie et al. 2009; Jolliffe and Cadima 2016); we adopt the simple and efficient profile likelihood method (Zhu and Ghodsi 2006) to automatically identify "elbow", which is the cut-off between the signal dimensions and the noise dimensions in scree plot.

Chernoff analysis

To analytically measure the performance of algorithms for block recovery, we consider the notion of Chernoff information among other possible metrics. Chernoff information enjoys the advantages of being independent of the clustering procedure, i.e., it can be derived no matter which clustering methods are used, and it is intrinsically relating to the Bayes risk (Tang and Priebe 2018; Athreya et al. 2017; Karrer and Newman 2011).

Definition 4 (*Chernoff Information* Chernoff 1952, 1956) Let F_1 and F_2 be two continuous multivariate distributions on $\mathbb{R}^d$ with density functions f_1 and f_2. The Chernoff information is defined as

$$\begin{aligned} C(F_1, F_2) &= -\log \left[\inf_{t\in(0,1)} \int_{\mathbb{R}^d} f_1^t(\mathbf{x}) f_2^{1-t}(\mathbf{x}) d\mathbf{x} \right] \\ &= \sup_{t\in(0,1)} \left[-\log \int_{\mathbb{R}^d} f_1^t(\mathbf{x}) f_2^{1-t}(\mathbf{x}) d\mathbf{x} \right]. \end{aligned} \tag{3}$$

Remark 3

Consider the special case where we take $F_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $F_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$*; then the corresponding Chernoff information is*

$$C(F_1, F_2) = \sup_{t\in(0,1)} \left[\frac{1}{2}t(1-t)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_t|}{|\boldsymbol{\Sigma}_1|^t|\boldsymbol{\Sigma}_2|^{1-t}}\right], \quad (4)$$

where $\boldsymbol{\Sigma}_t = t\boldsymbol{\Sigma}_1 + (1-t)\boldsymbol{\Sigma}_2$.

The comparsion of block recovery via Chernoff information is based on the statistical information between the limiting distributions of the blocks and smaller statistical information implies less information to discriminate between different blocks of the SBM. To that end, we also review the limiting results of ASE for SBM, essential for investigating Chernoff information.

Theorem 1 (*CLT of ASE for SBM* Rubin-Delanchy et al. 2022) *Let* $(\mathbf{A}^{(n)}, \mathbf{X}^{(n)}) \sim \text{GRDPG}(n, F, d_+, d_-)$ *be a sequence of adjacency matrices and associated latent positions of a d-dimensional GRDPG as in Definition* 1 *from an inner product distribution F where F is a mixture of K point masses in* $\mathbb{R}^d$*, i.e.,*

$$F = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\nu}_k} \qquad \text{with} \qquad \forall k,\ \pi_k > 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1, \quad (5)$$

where $\delta_{\boldsymbol{\nu}_k}$ *is the Dirac delta measure at* ν_k*. Let* $\Phi(\mathbf{z}, \boldsymbol{\Sigma})$ *denote the cumulative distribution function* (*CDF*) *of a multivariate Gaussian distribution with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}$*, evaluated at* $\mathbf{z} \in \mathbb{R}^d$*. Let* $\widehat{\mathbf{X}}^{(n)}$ *be the ASE of* $\mathbf{A}^{(n)}$ *with* $\widehat{\mathbf{X}}_i^{(n)}$ *as the i-th row* (*same for* $\mathbf{X}_i^{(n)}$)*. Then there exists a sequence of matrices* $\mathbf{M}_n \in \mathbb{R}^{d\times d}$ *satisfying* $\mathbf{M}_n \mathbf{I}_{d_+d_-} \mathbf{M}_n^\top = \mathbf{I}_{d_+d_-}$ *such that for all* $\mathbf{z} \in \mathbb{R}^d$ *and fixed index i,*

$$\mathbb{P}\left\{\sqrt{n}\left(\mathbf{M}_n \widehat{\mathbf{X}}_i^{(n)} - \mathbf{X}_i^{(n)}\right) \le \mathbf{z} \mid \mathbf{X}_i^{(n)} = \boldsymbol{\nu}_k\right\} \to \Phi(\mathbf{z}, \boldsymbol{\Sigma}_k), \quad (6)$$

where for $\boldsymbol{\nu} \sim F$

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}(\boldsymbol{\nu}_k) = \mathbf{I}_{d_+d_-} \boldsymbol{\Delta}^{-1} \mathbb{E}\left[\left(\boldsymbol{\nu}_k^\top \mathbf{I}_{d_+d_-} \boldsymbol{\nu}\right)\left(1 - \boldsymbol{\nu}_k^\top \mathbf{I}_{d_+d_-} \boldsymbol{\nu}\right)\boldsymbol{\nu}\boldsymbol{\nu}^\top\right] \boldsymbol{\Delta}^{-1} \mathbf{I}_{d_+d_-}, \quad (7)$$

with

$$\boldsymbol{\Delta} = \mathbb{E}\left[\boldsymbol{\nu}\boldsymbol{\nu}^\top\right]. \quad (8)$$

For a K-block SBM, let $\mathbf{B} \in (0,1)^{K\times K}$ be the block connectivity probability matrix and $\boldsymbol{\pi} \in (0,1)^K$ be the vector of block assignment probabilities. Given an n vertex instantiation of the SBM parameterized by $\mathbf{B}$ and $\boldsymbol{\pi}$, for sufficiently large n, the large sample optimal error rate for estimating the block assignments using ASE can be measured via Chernoff information as (Tang and Priebe 2018; Athreya et al. 2017)

$$\rho = \min_{k\neq l} \sup_{t\in(0,1)} \left[\frac{1}{2}nt(1-t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_{k\ell}(t)|}{|\boldsymbol{\Sigma}_k|^t|\boldsymbol{\Sigma}_\ell|^{1-t}}\right], \quad (9)$$

where $\mathbf{\Sigma}_{k\ell}(t) = t\mathbf{\Sigma}_k + (1-t)\mathbf{\Sigma}_\ell$, $\mathbf{\Sigma}_k = \mathbf{\Sigma}(\boldsymbol{\nu}_k)$ and $\mathbf{\Sigma}_\ell = \mathbf{\Sigma}(\boldsymbol{\nu}_\ell)$ are defined as in Eq. (7). Also note that as $n \to \infty$, the logarithm term in Eq. (9) will be dominated by the other term. Then we have the approximate Chernoff information as

$$\rho \approx \min_{k \neq l} C_{k,\ell}(\mathbf{B}, \boldsymbol{\pi}), \tag{10}$$

where

$$C_{k,\ell}(\mathbf{B}, \boldsymbol{\pi}) = \sup_{t \in (0,1)} \left[t(1-t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \mathbf{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) \right]. \tag{11}$$

We also introduce the following two notions, which will be used when we describe our dynamic network sampling scheme.

Definition 5 (*Chernoff-active Blocks*) For K-block SBM parametrized by the block connectivity probability matrix $\mathbf{B} \in (0,1)^{K \times K}$ and the vector of block assignment probabilities $\boldsymbol{\pi} \in (0,1)^K$. The Chernoff-active blocks (k^*, ℓ^*) are defined as

$$(k^*, \ell^*) = \arg\min_{k \neq l} C_{k,\ell}(\mathbf{B}, \boldsymbol{\pi}), \tag{12}$$

where $C_{k,\ell}(\mathbf{B}, \boldsymbol{\pi})$ is defined as in Eq. (10).

Definition 6 (*Chernoff Superiority*) For K-block SBMs, given two block connectivity probability matrices $\mathbf{B}, \mathbf{B}' \in (0,1)^{K \times K}$ and a vector of block assignment probabilities $\boldsymbol{\pi} \in (0,1)^K$. Let ρ_B and $\rho_{B'}$ denote the Chernoff information obtained as in Eq. (10) corresponding to $\mathbf{B}$ and $\mathbf{B}'$ respectively. We say that $\mathbf{B}$ is Chernoff superior to $\mathbf{B}'$, denoted as $\mathbf{B} \succ \mathbf{B}'$, if $\rho_B > \rho_{B'}$.

Remark 4

If $\mathbf{B}$ *is Chernoff superior to* $\mathbf{B}'$*, then we can have a better block recovery from* $\mathbf{B}$ *than* $\mathbf{B}'$*. In addition, Chernoff superiority is transitive, which is straightforward from the definition.*

Dynamic network sampling

We start our analysis with the unobserved block connectivity probability matrix $\mathbf{B}$ for SBM and then illustrate how to migrate the proposed methods for real applications when we have the observed adjacency matrix $\mathbf{A}$.

Consider the K-block SBM parametrized by the block connectivity probability matrix $\mathbf{B} \in (0,1)^{K \times K}$ and the vector of block assignment probabilities $\boldsymbol{\pi} \in (0,1)^K$ with $K > 2$. Given initial sampling parameter $p_0 \in (0,1)$, initial sampling is uniformly at random, i.e.,

$$\mathbf{B}_0 = p_0 \mathbf{B}. \tag{13}$$

This initial sampling simulates the case when one only obersves a partial graph with a small portion of the edges instead of the entire graph with all existing edges.

Theorem 2 *For K-block SBMs, given two block connectivity probability matrices* $\mathbf{B}, p\mathbf{B} \in (0,1)^{K \times K}$ *with* $p \in (0,1)$ *and a vector of block assignment probabilities* $\boldsymbol{\pi} \in (0,1)^K$*, we have* $\mathbf{B} \succ p\mathbf{B}$.

The proof of Theorem 2 can be found in Appendix. As an illustration, consider a 4-block SBM parametrized by block connectivity probability matrix $\mathbf{B}$ as

$$\mathbf{B} = \begin{bmatrix} 0.04 & 0.08 & 0.10 & 0.18 \\ 0.08 & 0.16 & 0.20 & 0.36 \\ 0.10 & 0.20 & 0.25 & 0.45 \\ 0.18 & 0.36 & 0.45 & 0.81 \end{bmatrix}. \tag{14}$$

Figure 1 shows Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14) and $p\mathbf{B}$ for $p \in (0,1)$. In addition, Fig. 1a assumes $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and Fig. 1b assumes $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$. As suggested by Theorem 2, for any $p \in (0,1)$ we have $\rho_B > \rho_{pB}$ and thus $\mathbf{B} \succ p\mathbf{B}$.

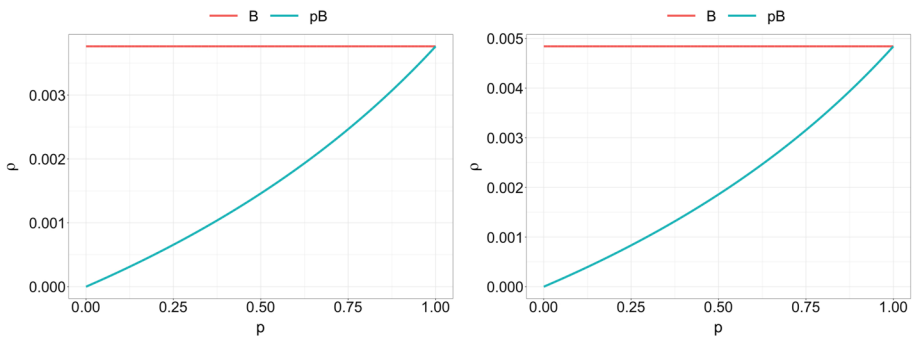
Now given dynamic network sampling parameter $p_1 \in (0, 1-p_0)$, the baseline sampling scheme can proceed uniformly at random again, i.e.,

$$\mathbf{B}_1 = \mathbf{B}_0 + p_1\mathbf{B} = (p_0 + p_1)\mathbf{B}. \tag{15}$$

This dynamic network sampling simulates the situation when one is given limited resources to sample some extra edges after observing the partial graph with only a small portion of the edges. Since we only have limited budget to sample another small portion of edges, one would benefit from identifying vertex pairs that have much influence on the community structure. In other words, the baseline sampling scheme just randomly choosing vertex pairs without using the information from the initial observed graphs and our goal is to design an alternative scheme to optimize this dynamic network sampling procedure so that one could have a better block recovery even with limited resources to only observe a partial graph with a small portion of the edges.

Corollary 1 *For K-block SBMs, given block connectivity probability matrix* $\mathbf{B} \in (0,1)^{K\times K}$ *and a vector of block assignment probabilities* $\boldsymbol{\pi} \in (0,1)^K$. *We have* $\mathbf{B} \succ \mathbf{B}_1 \succ \mathbf{B}_0$ *where* $\mathbf{B}_0$ *is defined as in Eq.* (13) *with* $p_0 \in (0,1)$ *and* $\mathbf{B}_1$ *is defined as in Eq.* (15) *with* $p_1 \in (0, 1-p_0)$.

The proof of Corollary 1 can be found in Appendix. This corollay implies that we can have a better block recovery from $\mathbf{B}_1$ than $\mathbf{B}_0$.



(a) balanced: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (b) unbalanced: $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$

Fig. 1 Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14) and $p\mathbf{B}$ for $p \in (0,1)$

Assumption 1 The Chernoff-active blocks after initial sampling is unique, i.e., there exists an unique pair $\left(k_0^*, \ell_0^*\right) \in \{(k,\ell) \mid 1 \le k < \ell \le K\}$ such that

$$\left(k_0^*, \ell_0^*\right) = \arg\min_{k \neq l} C_{k,\ell}(\mathbf{B}_0, \boldsymbol{\pi}), \tag{16}$$

where $\mathbf{B}_0$ is defined as in Eq. (13) and $\boldsymbol{\pi}$ is the vector of block assignment probabilities.

To improve this baseline sampling scheme, we concentrate on the Chernoff-active blocks $\left(k_0^*, \ell_0^*\right)$ after initial sampling assuming Assumption 1 holds. Instead of sampling from the entire block connectivity probability matrix $\mathbf{B}$ like the baseline sampling scheme as in Eq. (15), we only sample the entries associated with the Chernoff-active blocks. As a competitor to $\mathbf{B}_1$, our Chernoff-optimal dynamic network sampling scheme is then given by

$$\widetilde{\mathbf{B}}_1 = \mathbf{B}_0 + \frac{p_1}{\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2} \mathbf{B} \circ \mathbf{1}_{k_0^*, \ell_0^*}, \tag{17}$$

where $\circ$ denotes Hadamard product, $\pi_{k_0^*}$ and $\pi_{\ell_0^*}$ denote the block assignment probabilities for block k_0^* and ℓ_0^* respectively, and $\mathbf{1}_*$ is the $K \times K$ binary matrix with 0's everywhere except for 1's associated with the Chernoff-active blocks $\left(k_0^*, \ell_0^*\right)$, i.e., for any $i, j \in \{1, \cdots, K\}$

$$\mathbf{1}_{k_0^*, \ell_0^*}[i,j] = \begin{cases} 1 \text{ if } (i,j) \in \left\{\left(k_0^*, k_0^*\right),\ \left(k_0^*, \ell_0^*\right),\ \left(\ell_0^*, k_0^*\right),\ \left(\ell_0^*, \ell_0^*\right)\right\} \\ 0 \text{ otherwise} \end{cases}. \tag{18}$$

Note that the multiplier $\frac{1}{\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2}$ on $p_1 \mathbf{B} \circ \mathbf{1}_*$ assures that we sample the same number of potential edges with $\widetilde{\mathbf{B}}_1$ as we do with $\mathbf{B}_1$ in the baseline sampling scheme. In addition, to avoid over-sampling with respect to $\mathbf{B}$, i.e., to ensure $\widetilde{\mathbf{B}}_1[i,j] \le \mathbf{B}[i,j]$ for any $i, j \in \{1, \cdots, K\}$, we require

$$p_1 \le p_1^{\max} = (1 - p_0)\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2. \tag{19}$$

Assumption 2 For K-block SBMs, given a block connectivity probability matrix $\mathbf{B} \in (0,1)^{K \times K}$ and a vector of block assignment probabilities $\boldsymbol{\pi} \in (0,1)^K$. Let $p_1^* \in (0, p_1^{\max}]$ be the smallest positive $p_1 \le p_1^{\max}$ such that

$$\arg\min_{k \neq l} C_{k,\ell}(\widetilde{\mathbf{B}}_1, \boldsymbol{\pi}) \tag{20}$$

is not unique where $p_1^{\max}$ is defined as in Eq. (19) and $\widetilde{\mathbf{B}}_1$ is defined as in Eq. (17). If the arg min is always unique, let $p_1^* = p_1^{\max}$.

For any $p_1 \in (0, p_1^*)$, we can have a better block recovery from $\widetilde{\mathbf{B}}_1$ than $\mathbf{B}_1$, i.e., our Chernoff-optimal dynamic network sampling sheme is better than the baseline sampling scheme in terms of block recovery.

As an illustaration, consider the 4-block SBM with initial sampling parameter $p_0 = 0.01$ and block connectivity probability matrix $\mathbf{B}$ as in Eq. (14). Figure 2 shows the Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14), $\mathbf{B}_0$ as in Eq. (13), $\mathbf{B}_1$ as in Eq. (15), and $\widetilde{\mathbf{B}}_1$ as in Eq. (17) with dynamic network sampling parameter $p_1 \in (0, p_1^*)$ where p_1^* is defined as in Assumption 2. In addition, Figure 2a assumes $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and Fig. 2b assumes $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$. Note that for any $p_1 \in (0, p_1^*)$ we have $\rho_B > \rho_{\widetilde{B}_1} > \rho_{B_1} > \rho_{B_0}$ and thus $\mathbf{B} \succ \widetilde{\mathbf{B}}_1 \succ \mathbf{B}_1 \succ \mathbf{B}_0$. That is, in terms of Chernoff information, when given same amount of resources, the proposed Chernoff-optimal dynamic network sampling scheme can yield better block recovery results. In other words, to reach the same level of performance, in terms of Chernoff information, the proposed Chernoff-optimal dynamic network sampling scheme needs less resources.

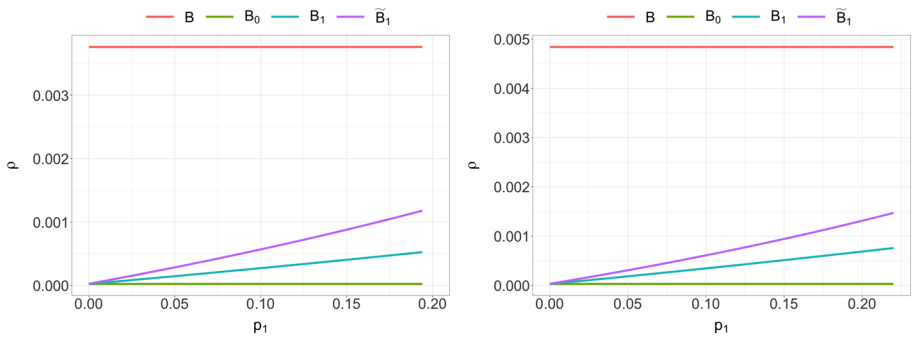
As described earlier, it may be the case that $p_1^* < p_1^{\max}$ at which point Chernoff-active blocks change to (k_1^*, ℓ_1^*). This potential non-uniquess of the Chernoff argmin is a consequence of our dynamic network sampling scheme. In the case of $p_1 > p_1^*$, our Chernoff-optimal dynamic network sampling scheme is adopted as

$$\widetilde{\mathbf{B}}_1^* = \mathbf{B}_0 + (p_1 - p_1^*)\mathbf{B} + \frac{p_1^*}{\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2}\mathbf{B} \circ \mathbf{1}_{k_0^*, \ell_0^*}, \tag{21}$$

Similarly, the multiplier $\frac{1}{\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2}$ on $p_1^*\mathbf{B} \circ \mathbf{1}_{k_0^*, \ell_0^*}$ assures that we sample the same number of potential edges with $\widetilde{\mathbf{B}}_1^*$ as we do with $\mathbf{B}_1$ in the baseline sampling scheme. In addition, to avoid over-sampling with respect to $\mathbf{B}$, i.e., $\widetilde{\mathbf{B}}_1^*[i,j] \le \mathbf{B}[i,j]$ for any $i, j \in \{1, \cdots, K\}$, we require

$$p_1 \le p_{11}^{\max} = 1 - p_0 - \frac{p_1^*}{\left(\pi_{k_0^*} + \pi_{\ell_0^*}\right)^2} + p_1^*. \tag{22}$$

For any $p_1 \in [p_1^*, p_{11}^{\max}]$, we can have a better block recovery from $\widetilde{\mathbf{B}}_1^*$ than $\mathbf{B}_1$, i.e., our Chernoff-optimal dynamic network sampling sheme is again better than the baseline sampling scheme in terms of block recovery.



(a) balanced: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (b) unbalanced: $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$

Fig. 2 Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14), $\mathbf{B}_0$ as in Eq. (13), $\mathbf{B}_1$ as in Eq. (15), and $\widetilde{\mathbf{B}}_1$ as in Eq. (17) with initial sampling parameter $p_0 = 0.01$ and dynamic network sampling parameter $p_1 \in (0, p_1^*)$ where p_1^* is defined as in Assumption 2

As an illustration, consider a 4-block SBM with initial sampling parameter $p_0 = 0.01$ and block connectivity probability matrix $\mathbf{B}$ as in Eq. (14). Figure 3 shows the Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14), $\mathbf{B}_0$ as in Eq. (13), $\mathbf{B}_1$ as in Eq. (15), and $\widetilde{\mathbf{B}}_1^*$ as in Eq. (21) with dynamic network sampling parameter $p_1 \in [p_1^*, p_{11}^{\max}]$ where p_1^* is defined as in Assumption 2 and $p_{11}^{\max}$ is defined as in Eq. (22). In addition, Fig. 3a assumes $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and Fig. 3b assumes $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$. Note that for any $p_1 \in [p_1^*, p_{11}^{\max}]$ we have $\rho_B > \rho_{\widetilde{B}_1^*} > \rho_{B_1} > \rho_{B_0}$ and thus $\mathbf{B} \succ \widetilde{\mathbf{B}}_1^* \succ \mathbf{B}_1 \succ \mathbf{B}_0$. That is, the adopted Chernoff-optimal dynamic network sampling scheme can still yield better block recovery results, in terms of Chernoff information, given the same amout of resources.

Now we illustrate how the proposed Chernoff-optimal dynamic network sampling sheme can be migrated for real applications. We summarize the uniform dynamic sampling scheme (baseline) as Algorithm 1 and our Chernoff-optimal dynamic network sampling scheme as Algorithm 2. Recall given potential edge set E and initial sampling parameter $p_0 \in (0, 1)$, we have the initial edge set $E_0 \subset E$ with $|E_0| = p_0|E|$. The goal is to dynamically sample new edges from the potential edge set so that we can have a better block recovery given limited resources.

Algorithm 1: Uniform dynamic network sampling scheme (baseline)

Input: Number of vertices n; potential edge set $E = \{(i,j) \mid i,j \in \{1,\cdots,n\}\}$; initial edge set $E_0 \subset E$; dynamic network sampling parameter $p_1 \in \left(0, 1 - \frac{|E_0|}{|E|}\right)$

1 Construct dynamic edge set as

$$E_1 = \{(i,j) \mid (i,j) \in E \setminus E_0\} \qquad \text{with} \qquad |E_1| = p_1|E|.$$

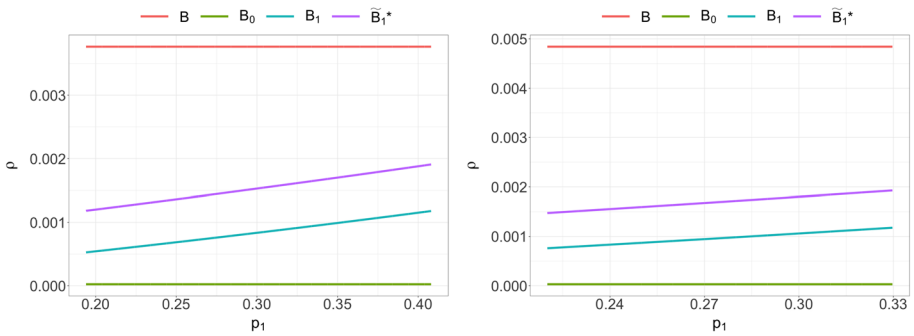
2 Construct dynamic adjacency matrix as $\mathbf{A} \in \{0,1\}^{n\times n}$ where for any $i,j \in \{1,\cdots,n\}$

$$\mathbf{A}[i,j] = \begin{cases} 1 & \text{if } (i,j) \in E_0 \bigcup E_1 \text{ or } (j,i) \in E_0 \bigcup E_1 \\ 0 & \text{otherwise} \end{cases}.$$

3 Estimate dynamic latent positions as $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times \widehat{d}}$ using ASE of $\mathbf{A}$ where $\widehat{d}$ is chosen as in Remark 2.

4 Cluster $\widehat{\mathbf{X}}$ using Gaussian mixture modeling (GMM) to estimate the block assignments as $\widehat{\boldsymbol{\tau}} \in \{1,\cdots,\widehat{K}\}^n$ where $\widehat{K}$ is chosen via Bayesian Information Criterion (BIC).

Output: Block assignments $\widehat{\boldsymbol{\tau}}$.



(a) balanced: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (b) unbalanced: $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$

Fig. 3 Chernoff information ρ as in Eq. (10) corresponding to $\mathbf{B}$ as in Eq. (14), $\mathbf{B}_0$ as in Eq. (13), $\mathbf{B}_1$ as in Eq. (15), and $\widetilde{\mathbf{B}}_1^*$ as in Eq. (21) with initial sampling parameter $p_0 = 0.01$ and dynamic network sampling parameter $p_1 \in [p_1^*, p_{11}^{\max}]$ where p_1^* is defined as in Assumption 2 and $p_{11}^{\max}$ is defined as in Eq. (22)

Algorithm 2: Chernoff-optimal dynamic network sampling scheme

Input: Number of vertices n; potential edge set $E = \{(i,j) \mid i,j \in \{1,\cdots,n\}\}$; initial edge set $E_0 \subset E$; dynamic network sampling parameter $p_1 \in \left(0, 1 - \frac{|E_0|}{|E|}\right)$

1 Construct dynamic adjacency matrix as $\mathbf{A} \in \{0,1\}^{n\times n}$ where for any $i,j \in \{1,\cdots,n\}$

$$\mathbf{A}[i,j] = \begin{cases} 1 & \text{if } (i,j) \in E_0 \text{ or } (j,i) \in E_0 \\ 0 & \text{otherwise} \end{cases}.$$

2 Estimate dynamic latent positions as $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times \widehat{d}}$ using ASE of $\mathbf{A}$ where $\widehat{d}$ is chosen as in Remark 2.

3 Cluster $\widehat{\mathbf{X}}$ using GMM to estimate the initial block assignments as $\widehat{\boldsymbol{\xi}} \in \{1,\cdots,\widehat{K}\}^n$ where $\widehat{K}$ is chosen via BIC.

4 Estimate the dynamic block assignment probability vector as $\widehat{\boldsymbol{\pi}} \in (0,1)^K$ where for $k \in \{1,\cdots,K\}$

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\widehat{\boldsymbol{\xi}}_i = k\}.$$

5 Estimate the dynamic block connectivity probability matrix as

$$\widehat{\mathbf{B}} = \widehat{\boldsymbol{\mu}}\mathbf{I}_{d_+ d_-}\widehat{\boldsymbol{\mu}}^\top \in [0,1]^{\widehat{K}\times\widehat{K}},$$

where $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^{\widehat{K}\times\widehat{d}}$ is the estimated means of all clusters.

6 Find the Chernoff-active blocks as

$$(k^*,\ell^*) = \arg\min_{k\neq l} C_{k,\ell}\left(\widehat{\mathbf{B}}, \widehat{\boldsymbol{\pi}}\right).$$

7 Construct dynamic edge set as

$$E_1 \subseteq E_* \quad \text{with} \quad |E_1| = \min\left\{p_1|E|\left(\widehat{\pi}_{k^*} + \widehat{\pi}_{\ell^*}\right)^2, |E_*|\right\},$$
$$E_{11} \subset E \setminus \left(E_0 \bigcup E_1\right) \quad \text{with} \quad |E_{11}| = p_1|E| - |E_1|,$$

where

$$E_* = \left\{(i,j) \mid (i,j) \in E \setminus E_0 \text{ and } \widehat{\xi}_i, \widehat{\xi}_j \in \{k^*,\ell^*\}\right\}.$$

8 Update dynamic adjacency matrix as $\mathbf{A} \in \{0,1\}^{n\times n}$ where for any $i,j \in \{1,\cdots,n\}$

$$\mathbf{A}[i,j] = \begin{cases} 1 & \text{if } (i,j) \in E_0 \bigcup E_1 \bigcup E_{11} \text{ or } (j,i) \in E_0 \bigcup E_1 \bigcup E_{11} \\ 0 & \text{otherwise} \end{cases}.$$

9 Update dynamic latent positions as $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times\widehat{d}}$ using ASE of updated $\mathbf{A}$ where $\widehat{d}$ is chosen as in Remark 2.

10 Cluster $\widehat{\mathbf{X}}$ using GMM to estimate the block assignments as $\widehat{\boldsymbol{\tau}} \in \{1,\cdots,\widehat{K}\}^n$ where $\widehat{K}$ is chosen via BIC.

Output: Block assignments $\widehat{\boldsymbol{\tau}}$.

Experiments

Simulations

In addition to Chernoff analysis, we also evalute our Chernoff-optimal dynamic network sampling sheme via simulations. In particular, consider the 4-block SBM parameterized by block connectivity probability matrix $\mathbf{B}$ as in Eq. (14) and dynamic network sampling parameter $p_1 \in (0, p_{11}^{\max}]$ where $p_{11}^{\max}$ is defined as in Eq. (22). We fix initial sampling parameter $p_0 = 0.01$. For each $p_1 \in (0, p_1^*)$ where p_1^* is defined as in Assumption 2, we simulate 50 adjacency matrices with $n = 12000$ vertices from $\mathbf{B}_1$ as in Eq. (15) and $\widetilde{\mathbf{B}}_1$ as in Eq. (17) respectively. For each $p_1 \in [p_1^*, p_{11}^{\max}]$, we simulate 50 adjacency matrices with $n = 12000$ vertices from $\mathbf{B}_1$ as in Eq. (15) and $\widetilde{\mathbf{B}}_1^*$ as in Eq. (21) respectively. In addition, Fig. 4a assumes $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, i.e., 3000 vertices in each

block, and Fig. 4b assumes $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$, i.e., 1500 vertices in two of the blocks and 4500 vertices in the other two blocks. We then apply ASE $\circ$ GMM (Step 3 and 4 in Algorithm 1) to recover block assignments and adopt adjusted Rand index (ARI) to measure the performance. Figure 4 shows ARI (`mean±stderr`) associated with $\mathbf{B}_1$ for $p_1 \in (0, p_{11}^{\max}]$, $\widetilde{\mathbf{B}}_1$ for $p_1 \in (0, p_1^*)$, and $\widetilde{\mathbf{B}}_1^*$ for $p_1 \in [p_1^*, p_{11}^{\max}]$ where the dashed lines denote p_1^*. Note that we can have a better block recovery from $\widetilde{\mathbf{B}}_1$ and $\widetilde{\mathbf{B}}_1^*$ than $\mathbf{B}_1$, which argee with our results from Chernoff analysis.

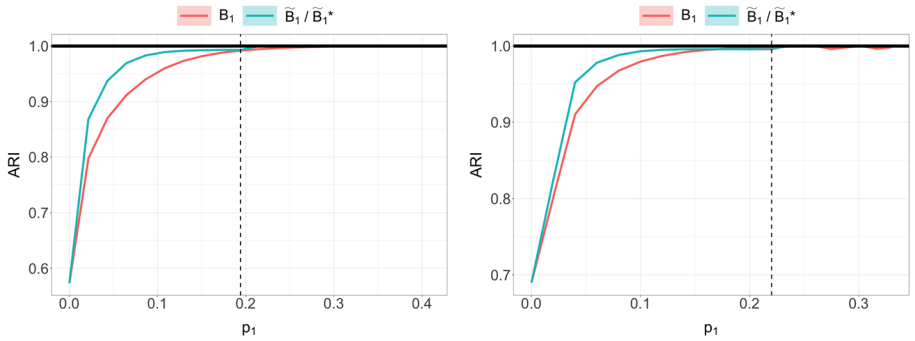
Now we compare the performance of Algorithms 1 and 2 by actual block recovery results. In particular, we start with the 4-block SBM parameterized by block connectivity probability matrix $\mathbf{B}$ as in Eq. (14). We consider dynamic network sampling parameter $p_1 \in (0, 1 - p_0)$ where p_0 is the initial sampling parameter. For each p_1, we simulate 50 adjacency matrices with $n = 4000$ vertices and retrieve associated potential edge sets. We fix initial sampling parameter $p_0 = 0.15$ and randomly sample initial edge sets. We then apply both algorithms to estimate the block assignments and adopt ARI to measure the performance. Figure 5 shows ARI (`mean±stderr`) of two algorithms for $p_1 \in (0, 0.85)$ where Fig. 5a assumes $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, i.e., 1000 vertices in each block, and Fig. 5b assumes $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$, i.e., 500 vertices in two of the blocks and 1500 vertices in the other two blocks. Note that both algorithms tend to have a better performance as p_1 increases, i.e., as we sample more edges, and Algorithm 2 can always recover more accurate block structure than Algorithm 1. That is, given the same amout of resources, the proposed Chernoff-optimal dynamic network sampling scheme can yield better block recovery results. In other words, to reach the same level of performance, in terms of the empirical clustering results, the proposed Chernoff-optimal dynamic network sampling scheme needs less resources.

Real data

We also evaluate the performance of Algorithms 1 and 2 for real application. We conduct real data experiments on a diffusion MRI connectome dataset (Priebe et al. 2019). There are 114 graphs (connectomes) estimated by the NDMG pipeline (Kiar et al. 2018) in this dataset. Each vertex in these graphs (the number of vertices n varies from 23728 to 42022) has a {Left, Right} hemisphere label and a {Gray, White} tissue label. We consider the potential 4 blocks as {LG, LW, RG, RW} where L and R denote the Left and Right hemisphere label, G and W denote the Gray and White tissue label. Here we consider initial sampling parameter $p_0 = 0.25$ and dynamic network sampling parameter $p_1 = 0.25$. Let $\Delta = \text{ARI(Algo2)} - \text{ARI(Algo1)}$ where ARI(Algo1) and ARI(Algo2) denotes the ARI when we apply Algorithms 1 and 2 respectively. The following hypothesis testing yields `p-value=0.0184`. Figure 6 shows algorithms' comparative performance via boxplot and histogram.

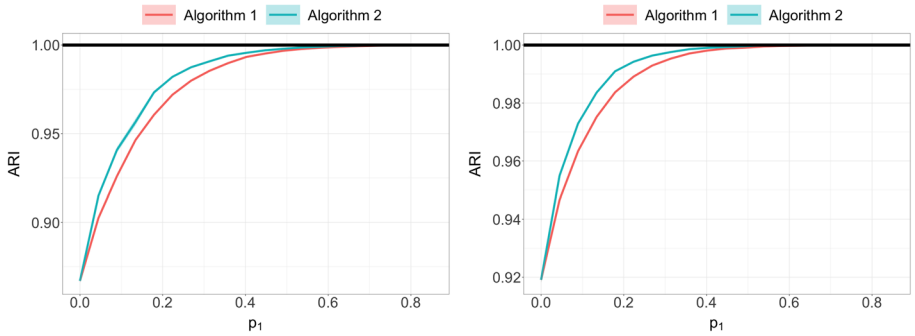
$$H_0 : \ \text{median}(\Delta) \leq 0 \qquad \text{v.s.} \qquad H_A : \ \text{median}(\Delta) > 0. \tag{23}$$

Furthermore, we test our algorithms on a Microsoft bing entity dataset (Agterberg et al. 2020). There are 2 graphs in this dataset where each has 13535 vertices. We treat block assignments estimated from the complete graph as ground truth. We consider initial sampling parameter $p_0 \in \{0.2, 0.3\}$ and dynamic network sampling parameter



(a) balanced: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (b) unbalanced: $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$

Fig. 4 Simulations for 4-block SBM parameterized by block connectivity probability matrix **B** as in Eq. (14) with initial sampling parameter $p_0 = 0.01$ and dynamic network sampling parameter $p_1 \in (0, p_{11}^{\max}]$ where $p_{11}^{\max}$ is defined as in Eq. (22). The dashed lines denote p_1^* which is defined as in Assumption 2



(a) balanced: $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (b) unbalanced: $\boldsymbol{\pi} = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$

Fig. 5 Simulations for 4-block SBM parameterized by block connectivity probability matrix **B** as in Eq. (14) with initial sampling parameter $p_0 = 0.15$ and dynamic network sampling parameter $p_1 \in (0, 0.85)$

$p_1 \in \{0,\ 0.05,\ 0.1,\ 0.15,\ 0.2\}$. For each p_1, we sample 100 times and compare the overall performance of Algorithm 1 and 2. Figure 7 shows the results where ARI is reported as `mean(±stderr)`.

We also conduct real data experiments with 2 social network datasets.

- LastFM asia social network data set (Leskovec and Krevl 2014; Rozemberczki and Sarkar 2020): Vertices (the number of vertices $n = 7624$) represent LastFM users from asian countries and edges (the number of edges $e = 27806$) represent mutual follower relationships. We treat 18 different location of users, which are derived from the country field for each user, as the potential block.
- Facebook large page-page network data set (Leskovec and Krevl 2014; Rozemberczki et al. 2019): Vertices (the number of vertices $n = 22470$) represent official Facebook pages and edges (the number of edges $e = 171002$) represent mutual likes. We treat 4 page types {Politician, Governmental Organization, Television Show, Company}, which are defined by Facebook, as the potential block.

We consider initial sampling parameter $p_0 \in \{0.15,\ 0.35\}$ and dynamic network sampling parameter $p_1 \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$. For each p_1, we sample 100 times and compare the overall performance of Algorithm 1 and 2. Figure 8 shows the results where ARI is reported as `mean(±stderr)`. Again it suggests that given the same amout of resources, the proposed Chernoff-optimal dynamic network sampling scheme can yield better block recovery results. In other words, to reach the same level of performance, in terms of the empirical clustering results, the proposed Chernoff-optimal dynamic network sampling scheme needs less resources.

Discussion

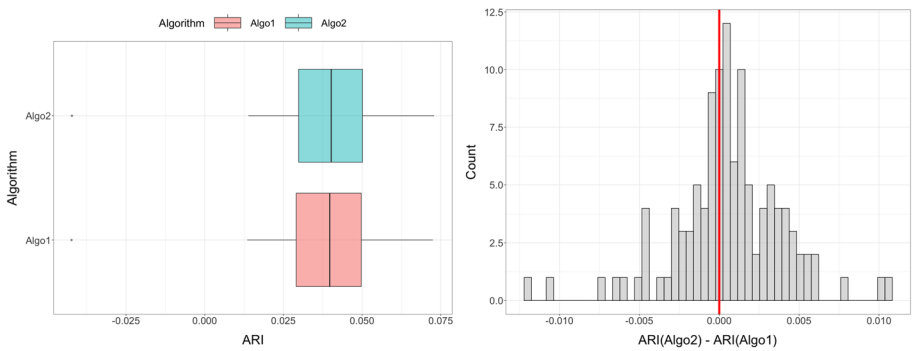
We propose a dynamic network sampling scheme to optimize block recovery for SBM when we only have a limited budget to observe a partial graph. Theoretically, we provide justification of our proposed Chernoff-optimal dynamic sampling scheme via the Chernoff information. Practically, we evaluate the performance, in terms of block recovery (community detection), of our method on several real datasets including diffusion MRI connectome dataset, Microsoft bing entity graph transitions dataset and social network datasets. Both theoretically and practically results suggest that our method can identify vertices that have the most impact on block structure and only check whether there are edges between them to save significant resources but still recover the block structure.

As the Chernoff-optimal dynamic sampling scheme depends on the initial clustering results to identify Chernoff-active blocks and construct dynamic edge set. Thus the performance could be impacted if the initial clustering is not very ideal. One of the future direction is to design certain strategy to reduce this dependency such that the proposed scheme is more robust.

Appendix

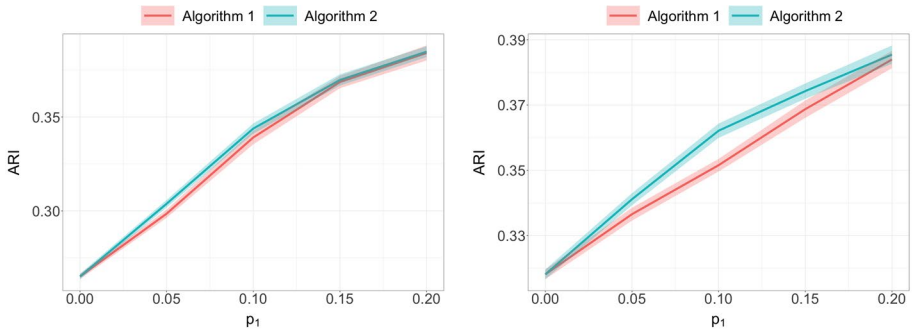
Proof of Theorem 2

Let $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{U}^\top$ *be the spectral decomposition of* $\mathbf{B}$ *and* $\mathbf{B}' = p\mathbf{B}$ *with* $p \in (0, 1)$. *Then we have*



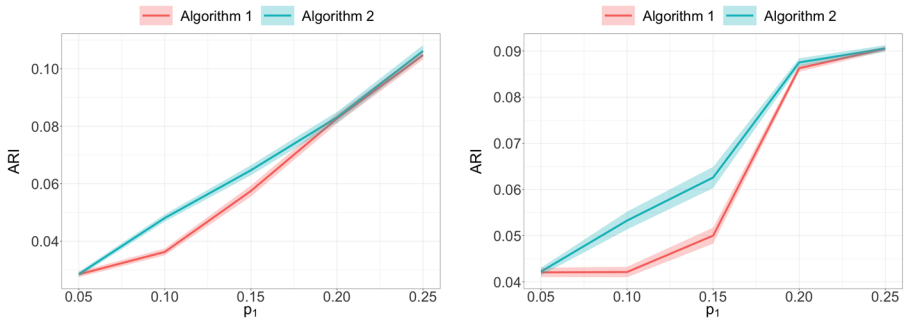
(a) Boxplot (b) Histogram

Fig. 6 Algorithms' comparative performance on diffusion MRI connectome data via ARI with initial sampling parameter $p_0 = 0.25$ and dynamic network sampling parameter $p_1 = 0.25$



(a) $p_0 = 0.2,\ p_1 \in \{0,\ 0.05,\ 0.1,\ 0.15,\ 0.2\}$ (b) $p_0 = 0.3,\ p_1 \in \{0,\ 0.05,\ 0.1,\ 0.15,\ 0.2\}$

Fig. 7 Algorithms' comparative performance on Microsoft bing entity data via ARI with different initial sampling parameter p_0 and dynamic network sampling parameter p_1



(a) LastFM: $p_0 = 0.15,\ p_1 \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$ (b) Facebook: $p_0 = 0.35,\ p_1 \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$

Fig. 8 Algorithms' comparative performance on social network data via ARI with different initial sampling parameter p_0 and dynamic network sampling parameter p_1

$$\mathbf{B}' = \mathbf{U}'\mathbf{S}\left(\mathbf{U}'\right)^\top \qquad \text{where} \qquad \mathbf{U}' = \sqrt{p}\mathbf{U}. \tag{24}$$

By Remark 1, to represent these two SBMs parametrized by two block connectivity matrices $\mathbf{B}$ and $\mathbf{B}'$ respectively (with the same block assignment probability vector $\boldsymbol{\pi}$) in the GRDPG models, we can take

$$\begin{aligned} \boldsymbol{v} &= \left[\boldsymbol{v}_1\ \cdots\ \boldsymbol{v}_K\right]^\top = \mathbf{U}|\mathbf{S}|^{1/2} \in \mathbb{R}^{K\times d}, \\ \boldsymbol{v}' &= \left[\boldsymbol{v}'_1\ \cdots\ \boldsymbol{v}'_K\right]^\top = \mathbf{U}'|\mathbf{S}|^{1/2} = \sqrt{p}\mathbf{U}|\mathbf{S}|^{1/2} = \sqrt{p}\boldsymbol{v} \in \mathbb{R}^{K\times d}. \end{aligned} \tag{25}$$

Then for any $k \in \{1, \cdots, K\}$, we have $\boldsymbol{v}'_k = \sqrt{p}\boldsymbol{v}_k \in \mathbb{R}^d$. By Theorem 1, we have

$$\begin{aligned} \boldsymbol{\Delta} &= \sum_{k=1}^{K} \pi_k \boldsymbol{v}_k \boldsymbol{v}_k^\top \in \mathbb{R}^{d\times d}, \\ \boldsymbol{\Delta}' &= \sum_{k=1}^{K} \pi_k \boldsymbol{v}'_k \left(\boldsymbol{v}'_k\right)^\top = p\sum_{k=1}^{K} \pi_k \boldsymbol{v}_k \boldsymbol{v}_k^\top = p\boldsymbol{\Delta} \in \mathbb{R}^{d\times d}. \end{aligned} \tag{26}$$

Note that $\mathbf{B}$ and $\mathbf{B}'$ have the same eigenvalues, thus we have $\mathbf{I}_{d_+d_-} = \mathbf{I}'_{d_+d_-}$. See also Lemma 2 (Gallagher et al. 2019). Then for $k \in \{1, \cdots, K\}$, we have

$$\begin{aligned}
\boldsymbol{\Sigma}_k &= \mathbf{I}_{d_+d_-}\boldsymbol{\Delta}^{-1}\mathbb{E}\Big[\Big(\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}\Big)\Big(1-\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}\Big)\boldsymbol{v}\boldsymbol{v}^\top\Big]\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \\
&= \mathbf{I}_{d_+d_-}\boldsymbol{\Delta}^{-1}\left[\sum_{\ell=1}^{K}\pi_\ell\Big(\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\Big(1-\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top\right]\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \in \mathbb{R}^{d\times d}, \\
[1em]\boldsymbol{\Sigma}'_k &= \frac{1}{p^2}\mathbf{I}_{d_+d_-}\boldsymbol{\Delta}^{-1}\left[p^2\sum_{\ell=1}^{K}\pi_\ell\Big(\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\Big(1-p\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top\right]\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \\
&= \mathbf{I}_{d_+d_-}\boldsymbol{\Delta}^{-1}\left[p\sum_{\ell=1}^{K}\pi_\ell\Big(\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\Big(1-\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top\right]\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \\
&\quad + \mathbf{I}_{d_+d_-}\boldsymbol{\Delta}^{-1}\left[(1-p)\sum_{\ell=1}^{K}\pi_\ell\Big(\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell\Big)\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top\right]\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \\
&= p\boldsymbol{\Sigma}_k + \mathbf{V}^\top\mathbf{D}_k(p)\mathbf{V} \in \mathbb{R}^{d\times d},
\end{aligned} \tag{27}$$

where

$$\begin{aligned}
\mathbf{V} &= \boldsymbol{v}\boldsymbol{\Delta}^{-1}\mathbf{I}_{d_+d_-} \in \mathbb{R}^{K\times d}, \\
\mathbf{D}_k(p) &= (1-p)\mathrm{diag}\Big(\pi_1\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_1, \cdots, \pi_K\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_K\Big) \in (0,1)^{K\times K}.
\end{aligned} \tag{28}$$

Recall that by Remark 1, we have $\boldsymbol{v}_k^\top \mathbf{I}_{d_+d_-}\boldsymbol{v}_\ell = \mathbf{B}_{k\ell} \in (0,1)$ for all $k, \ell \in \{1, \cdots, K\}$. Then we have $\mathbf{D}_k(p)$ is positive-definite for any $k \in \{1, \ldots, K\}$ and $p \in (0,1)$. For $k, \ell \in \{1, \ldots, K\}$ and $t \in (0,1)$, let $\boldsymbol{\Sigma}_{k\ell}(t)$ and $\boldsymbol{\Sigma}'_{k\ell}(t)$ denote the matrics as in Eq. (10) corresponding to $\mathbf{B}$ and $\mathbf{B}'$ respectively, i.e.,

$$\begin{aligned}
\boldsymbol{\Sigma}_{k\ell}(t) &= t\boldsymbol{\Sigma}_k + (1-t)\boldsymbol{\Sigma}_\ell \in \mathbb{R}^{d\times d}, \\
[1em]\boldsymbol{\Sigma}'_{k\ell}(t) &= t\boldsymbol{\Sigma}'_k + (1-t)\boldsymbol{\Sigma}'_\ell \\
&= t\Big[p\boldsymbol{\Sigma}_k + \mathbf{V}^\top\mathbf{D}_k(p)\mathbf{V}\Big] + (1-t)\Big[p\boldsymbol{\Sigma}_\ell + \mathbf{V}^\top\mathbf{D}_\ell(p)\mathbf{V}\Big] \\
&= p[t\boldsymbol{\Sigma}_k + (1-t)\boldsymbol{\Sigma}_\ell] + \mathbf{V}^\top[t\mathbf{D}_k(p) + (1-t)\mathbf{D}_\ell(p)]\mathbf{V} \\
&= p\boldsymbol{\Sigma}_{k\ell}(t) + \mathbf{V}^\top\mathbf{D}_{k\ell}(p,t)\mathbf{V} \in \mathbb{R}^{d\times d},
\end{aligned} \tag{29}$$

where

$$\mathbf{D}_{k\ell}(p,t) = t\mathbf{D}_k(p) + (1-t)\mathbf{D}_\ell(p) \in \mathbb{R}_+^{K\times K}. \tag{30}$$

Recall that $\mathbf{D}_k(p)$ and $\mathbf{D}_\ell(p)$ are both positive-definite for any $k, \ell \in \{1, \ldots, K\}$ and $p \in (0,1)$, thus $\mathbf{D}_{k\ell}(p,t)$ is also positive-definite for any $k, \ell \in \{1, \ldots, K\}$ and$p, t \in (0,1)$. Now by the Sherman-Morrison-Woodbury formula (Horn and Johnson 2012), we have

$$\begin{aligned}\left[\boldsymbol{\Sigma}'_{k\ell}(t)\right]^{-1} &= \left[p\boldsymbol{\Sigma}_{k\ell}(t) + \mathbf{V}^\top \mathbf{D}_{k\ell}(p,t)\mathbf{V}\right]^{-1} \\ &= \frac{1}{p}\boldsymbol{\Sigma}_{k\ell}^{-1}(t) - \frac{1}{p^2}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\mathbf{V}^\top \left[\mathbf{D}_{k\ell}^{-1}(p,t) + \frac{1}{p}\mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\mathbf{V}^\top\right]^{-1}\mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t) \\ &= \frac{1}{p}\boldsymbol{\Sigma}_{k\ell}^{-1}(t) - \frac{1}{p^2}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\mathbf{V}^\top \mathbf{M}_{k\ell}^{-1}(p,t)\mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t) \in \mathbb{R}^{d\times d},\end{aligned} \tag{31}$$

where

$$\mathbf{M}_{k\ell}(p,t) = \mathbf{D}_{k\ell}^{-1}(p,t) + \frac{1}{p}\mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\mathbf{V}^\top \in \mathbb{R}^{K\times K}. \tag{32}$$

Recall that for any $k,\ell \in \{1,\dots,K\}$ and $p,t \in (0,1)$, $\mathbf{D}_{k\ell}(p,t)$ and $\boldsymbol{\Sigma}_{k\ell}(t)$ are both positive-definite, thus $\mathbf{M}_{k\ell}(p,t)$ is also positive-definite. Then for any $k,\ell \in \{1,\dots,K\}$ and $p,t \in (0,1)$, we have

$$\begin{aligned}(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell)^\top \left[\boldsymbol{\Sigma}'_{k\ell}(t)\right]^{-1}(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell) &= p(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \\ &\quad \left[\frac{1}{p}\boldsymbol{\Sigma}_{k\ell}^{-1}(t) - \frac{1}{p^2}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\mathbf{V}^\top \mathbf{M}_{k\ell}^{-1}(p,t)\mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)\right] \\ &\quad (\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) \\ &= (\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) \\ &\quad - \frac{1}{p}\mathbf{x}^\top \mathbf{M}_{k\ell}^{-1}(p,t)\mathbf{x} \\ &= (\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) - h_{k\ell}(p,t),\end{aligned} \tag{33}$$

where

$$\begin{aligned}\mathbf{x} &= \mathbf{V}\boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) \in \mathbb{R}^K, \\ h_{k\ell}(p,t) &= \frac{1}{p}\mathbf{x}^\top \mathbf{M}_{k\ell}^{-1}(p,t)\mathbf{x}.\end{aligned} \tag{34}$$

Recall that for any $k,\ell \in \{1,\dots,K\}$ and $p,t \in (0,1)$, $\mathbf{M}_{k\ell}(p,t)$ is positive-definite, thus we have $h_{k\ell}(p,t) > 0$. Together with Eq. (33), we have

$$t(1-t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell) > t(1-t)(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell)^\top \left[\boldsymbol{\Sigma}'_{k\ell}(t)\right]^{-1}(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell). \tag{35}$$

Thus for any $k,\ell \in \{1,\dots,K\}$, we have

$$\begin{aligned}C_{k,\ell}(\mathbf{B},\boldsymbol{\pi}) &= \sup_{t\in(0,1)} \left[t(1-t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)^\top \boldsymbol{\Sigma}_{k\ell}^{-1}(t)(\boldsymbol{\nu}_k - \boldsymbol{\nu}_\ell)\right], \\ &> \sup_{t\in(0,1)} \left[t(1-t)(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell)^\top \left[\boldsymbol{\Sigma}'_{k\ell}(t)\right]^{-1}(\boldsymbol{\nu}'_k - \boldsymbol{\nu}'_\ell)\right] \\ &= C_{k,\ell}(\mathbf{B}',\boldsymbol{\pi}).\end{aligned} \tag{36}$$

Let ρ_B and $\rho_{B'}$ denote the Chernoff information obtained as in Eq. (10) corresponding to $\mathbf{B}$ and $\mathbf{B}'$ respectively (with the same block assignment probability vector $\boldsymbol{\pi}$). Then we have

$$\rho_B \approx \min_{k \neq l} C_{k,\ell}(\mathbf{B}, \boldsymbol{\pi}) > \min_{k \neq l} C_{k,\ell}(\mathbf{B}', \boldsymbol{\pi}) \approx \rho_{B'}. \tag{37}$$

Thus we have $\mathbf{B} \succ \mathbf{B}' = p\mathbf{B}$ for $p \in (0, 1)$. □

Proof of Corollary 1

By Eq. (13) and Eq. (15), we have

$$\begin{aligned} \mathbf{B}_0 &= \frac{p_0}{p_0 + p_1}\mathbf{B}_1, \\ \mathbf{B}_1 &= (p_0 + p_1)\mathbf{B}. \end{aligned} \tag{38}$$

Recall that $p_0 \in (0, 1)$ and $p_1 \in (0, 1 - p_0)$. Then by Theorem 2, we have $\mathbf{B} \succ \mathbf{B}_1 \succ \mathbf{B}_0$. □

Abbreviations

SBM	Stochastic Blockmodel
GRDPG	Generalized random dot product graph
ASE	Adjacency spectral embedding
LSE	Laplacian spectral embedding
GMM	Gaussian mixture modeling
BIC	Bayesian information criterion
ARI	Adjusted Rand index
stderr	Standard error
NDMG	NeuroData's magnetic resonance imaging to graphs

Acknowledgements
This problem was posed to us by Adam Cardinal-Stakenas and Kevin Hoover.

Author contributions
CM developed the theory, designed and implemented the methods, conducted the experiments, and wrote the manuscript . YP implemented the methods, conducted the experiments, and edited the manuscript. CEP formulated the problem, designed the methods, developed the theory and edited the manuscript. All authors read and approved the manuscript.

Funding
Cong Mu's work is partially supported by the Johns Hopkins Mathematical Institute for Data Science (MINDS) Data Science Fellowship.

Availibility of data and materials
Social network datasets are available at https://www.snap.stanford.edu/data/.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 31 August 2022 Accepted: 16 December 2022
Published online: 13 January 2023

References

Agterberg J, Park Y, Larson J, White C, Priebe CE, Lyzinski V (2020) Vertex nomination, consistent estimation, and adversarial modification. Electron J Stat 14(2):3230–3267
Athreya A, Priebe CE, Tang M, Lyzinski V, Marchette DJ, Sussman DL (2016) A limit theorem for scaled eigenvectors of random dot product graphs. Sankhya A 78(1):1–18
Athreya A, Fishkind DE, Tang M, Priebe CE, Park Y, Vogelstein JT, Levin K, Lyzinski V, Qin Y (2017) Statistical inference on random dot product graphs: a survey. J Mach Learn Res 18(1):8393–8484
Binkiewicz N, Vogelstein JT, Rohe K (2017) Covariate-assisted spectral clustering. Biometrika 104(2):361–377

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 23(4):493–507
Chernoff H (1956) Large-sample theory: parametric case. Ann Math Stat 27(1):1–22
Choi DS, Wolfe PJ, Airoldi EM (2012) Stochastic blockmodels with a growing number of classes. Biometrika 99(2):273–284
Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44
Gallagher I, Bertiger A, Priebe C, Rubin-Delanchy P (2019) Spectral clustering in the weighted stochastic block model. arXiv:1910.05534
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5(2):109–137
Horn RA, Johnson CR (2012) Matrix Analysis. Cambridge University Press, New York
Huang S, Feng Y (2018) Pairwise covariates-adjusted block model for community detection. arXiv:1807.03469
Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci 374(2065):20150202
Karrer B, Newman ME (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83(1):016107
Kiar G, Bridgeford EW, Gray Roncal WR, Chandrashekhar V, Mhembere D, Ryman S, Zuo X-N, Margulies DS, Craddock RC, Priebe CE, Jung R, Calhoun VD, Caffo B, Burns R, Milham MP, Vogelstein JT (2018) A high-throughput pipeline identifies robust connectomes but troublesome variability. bioRxiv, 188706
Leskovec J, Krevl A (2014) SNAP datasets: stanford large network dataset collection. http://snap.stanford.edu/data
Lyzinski V, Sussman DL, Tang M, Athreya A, Priebe CE (2014) Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. Electron J Stat 8(2):2905–2922
Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE (2016) Community detection and classification in hierarchical stochastic blockmodels. IEEE Trans Netw Sci Eng 4(1):13–26
McSherry F (2001) Spectral partitioning of random graphs. In: Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pp 529–537. IEEE
Mele A, Hao L, Cape J, Priebe CE (2022) Spectral inference for large stochastic blockmodels with nodal covariates. J Bus Econ Stat
Mu C, Mele A, Hao L, Cape J, Athreya A, Priebe CE (2022) On spectral algorithms for community detection in stochastic blockmodel graphs with vertex covariates. IEEE Trans Netw Sci Eng
Priebe CE, Park Y, Vogelstein JT, Conroy JM, Lyzinski V, Tang M, Athreya A, Cape J, Bridgeford E (2019) On a two-truths phenomenon in spectral graph clustering. Proc Natl Acad Sci 116(13):5995–6000
Purohit S, Choudhury S, Holder LB (2017) Application-specific graph sampling for frequent subgraph mining and community detection. In: 2017 IEEE International Conference on Big Data (Big Data), pp 1000–1005. IEEE
Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. Ann Stat 39(4):1878–1915
Roy S, Atchadé Y, Michailidis G (2019) Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. J Comput Graph Stat 28(3):609–619
Rozemberczki B, Allen C, Sarkar R (2019) Multi-scale attributed node embedding. arXiv:1909.13021
Rozemberczki B, Sarkar R (2020) Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In: Proceedings of the 29th ACM International conference on information and knowledge management (CIKM '20), pp 1325–1334. ACM
Rubin-Delanchy P, Priebe CE, Tang M, Cape J (2022) A statistical interpretation of spectral embedding: the generalised random dot product graph. J R Stat Soc
Sussman DL, Tang M, Fishkind DE, Priebe CE (2012) A consistent adjacency spectral embedding for stochastic blockmodel graphs. J Am Stat Assoc 107(499):1119–1128
Sweet TM (2015) Incorporating covariates into stochastic blockmodels. J Educ Behav Stat 40(6):635–664
Tang M, Priebe CE (2018) Limit theorems for eigenvectors of the normalized laplacian for random graphs. Ann Stat 46(5):2360–2415
Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
Yun S-Y, Proutiere A (2014) Community detection via random and adaptive sampling. In: Conference on Learning Theory, pp 138–175. PMLR
Zhu M, Ghodsi A (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. Comput Stati Data Anal 51(2):918–930

Publisher's Note