Subsampling Based Community Detection for Large Networks

Sayan Chakrabarty[1], *Srijan Sengupta*[2] *and Yuguo Chen*[1]

[1]*University of Illinois at Urbana-Champaign*

[2]*North Carolina State University*

Supplementary Material

S1 Additional Definitions and Algorithms

S1.1 Label Matching

Here, we present two algorithms for matching two sets of labels. Algorithm S1 searches over all possible permutations of the first set of labels to select the permutation that has the least number of mismatched nodes with the second set of labels. Although this algorithm gives the best permutation of the labels, it is computationly not feasible for large number of communities. Algorithm S2 computes the number of nodes in each pair of communities between the two sets of labels. It swaps the communities between the two sets of labels that have the highest number of nodes between them. Algo-

rithm S2 is shown to produce the same permutation matrix as Algorithm S1 when one of the community membership matrices can be expressed as a permutation of the other plus some error term that satisfies certain conditions (Mukherjee et al., 2021).

Algorithm S1 `MatchBF`

Input Two community membership matrices C_1 and C_2 with the same number of communities

Output A permutation matrix P that best aligns C_1 with C_2

procedure `MATCHBF`(C_1, C_2)

$K \leftarrow$ number of columns of C_1

$\mathbf{E}_K \leftarrow$ list of all permutation matrices of order $K \times K$

initialize a vector $mismatch$ of length $K!$

$i \leftarrow 1$

for $E \in \mathbf{E}_K$ **do**

$mismatch[i] \leftarrow \|(C_1 E - C_2)\|_0$

$i \leftarrow i + 1$

return $\mathbf{E}_K[\arg\min_i mismatch[i]]$

Algorithm S2 `MatchGreedy`

Input Two community membership matrices C_1 and C_2 with the same number of communities

Output A permutation matrix P that approximates the best alignment of C_1 with C_2

procedure MATCHGREEDY(C_1, C_2)

$K \leftarrow$ number of columns of C_1

$P \leftarrow \mathbf{0}_{K \times K}$ (null matrix of order $K \times K$)

$M \leftarrow C_1^T C_2$

while there are rows or columns of M left with positive values **do**

find $(i, j) = \arg\max_{i,j} M_{ij}$ (ties are broken arbitrarily)

$P_{ij} \leftarrow 1$

replace the ith row and the jth column of M by -1

return P

S1.2 Spectral Clustering on SBM

Stochastic Blockmodel A stochastic blockmodel with n nodes and K communities is parameterized by a pair of matrices (C, P), where $C \in \mathbb{C}_{n \times K}$ is a membership matrix, and $P \in \mathbb{R}_{K \times K}$ is a symmetric connectivity matrix. For each node i, let g_i $(1 \leq g_i \leq K)$ be its community label such that the ith row of C is 1 in column g_i and 0 elsewhere. The entry P_{kl} in P is the edge probability between a node in community k and a node in community

l. Given (C, P), the adjacency matrix $A = (a_{ij})_{1 \le i,j \le n}$ is generated as

$$a_{ij} = \begin{cases} \text{independent Bernoulli}(P_{g_i g_j}), & \text{if } i < j, \\ 0, & \text{if } i = j, \\ a_{ji}, & \text{if } i > j. \end{cases} \tag{S1.1}$$

Also, let $n_1, \ldots, n_K$ be the size of each community in the SBM.

Spectral Clustering Spectral clustering is a simple community detection method that recovers the underlying community structures of a network using the eigen decomposition of the corresponding adjacency matrix. For an undirected simple network with adjacency matrix A, spectral clustering computes the eigenvectors and eigenvalues of A. Then it clusters the K eigenvectors corresponding to the largest K eigenvalues in terms of their absolute values. Any clustering algorithm can be used at this step. However, we consider K-means clustering at this step. Since solving a K-means clustering problem is NP-hard, we use $(1 + \delta)$-approximate solution to the K-means problem. Spectral clustering with approximate K-means is summarized in Algorithm S3.

Algorithm S3 Spectral Clustering with Approximate K-means

Input An adjacency matrix $A_{n\times n}$, number of communities K, and an approximating parameter $\delta > 0$ for K-means clustering.

Output A membership matrix $\hat{C}_{n\times K}$.

procedure SC(A, K)

1. Calculate $\hat{U} \in \mathbb{R}^{n\times K}$ consisting of the leading K eigenvectors (ordered in absolute eigenvalue) of A.

2. Let $(\hat{C}, \hat{X}) \in \mathbb{C}_{n\times K} \times \mathbb{R}^{K\times K}$ such that $\|\hat{C}\hat{X} - \hat{U}\|_F^2 \leq (1 + \delta) \min_{(C,X)\in\mathbb{C}_{n\times K}\times\mathbb{R}^{K\times K}} \|CX - \hat{U}\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm.

3. Output $\hat{C}$.

Theorem S1. *(Lei and Rinaldo, 2015) Let A be an adjacency matrix of a simple undirected network generated from a stochastic blockmodel SBM($C_{n\times K}$, $P_{K\times K}$). Assume that*

1. *$P = \alpha_n P_0$ for some $\alpha_n \geq \frac{1}{n}\log n$,*

2. *P_0 has minimum absolute eigenvalue $\geq \lambda > 0$,*

3. *$\max_{k,l} P_{0_{kl}} = 1$,*

4. *$\hat{C}$ is the solution of $(1 + \delta)$-approximate K-means spectral clustering (Algorithm S3) applied on A with K communities.*

Then there exists an absolute constant c such that if the parameters (n, K, α_n, λ, δ) *satisfy*

$$(2+\delta)\frac{K}{\lambda^2\pi^2_{min}n\alpha_n} < c, \tag{S1.2}$$

then with probability at least $1-\frac{1}{n}$,

$$\delta(\hat{C}, C) = \frac{1}{n}\min_{E\in\mathbf{E}_K}\|\hat{C}E - C\|_0 \le c^{-1}(2+\delta)\frac{K\pi_{max}}{\lambda^2\pi^2_{min}n\alpha_n}. \tag{S1.3}$$

S1.3 Spherical K-median Spectral Clustering on DCBM

Degree Corrected Blockmodel A degree corrected blockmodel with n nodes and K communities is parameterized by a triplet (C, P, ψ), where $C_{n\times K}$ and $P_{K\times K}$ are defined similarly as in SBM, and the vector $\psi \in \mathbb{R}^n$ is included to model additional variability of the edge probabilities at the node level. Given (C, P, ψ), the adjacency matrix $A = (a_{ij})_{1\le i,j\le n}$ is generated as

$$a_{ij} = \begin{cases} \text{independent Bernoulli}(\psi_i\psi_j P_{g_ig_j}), & \text{if } i < j, \\ 0, & \text{if } i = j, \\ a_{ji}, & \text{if } i > j. \end{cases} \tag{S1.4}$$

We assume $\max_{i\in G_k}\psi_i = 1$ to avoid the problem of identifiability, where G_k is the set of all nodes in the kth community, $1 \le k \le K$.

Spherical K-median Spectral Clustering Community recovery is difficult for a DCBM due to the presence of degree heterogeneity. Small values in

ψ makes it hard to identify the community membership of the corresponding nodes as few edges are observed for those nodes. Spherical K-median spectral clustering overcomes this issue by row-normalizing the top K eigenvector matrix $\hat{U}$ and then minimizing the matrix $2,1$ distance between the points and cluster centers. The results on DCBM discussed in Lei and Rinaldo (2015) were derived for spherical $(1+\delta)$-approximate K-median spectral clustering. The algorithm is presented in Algorithm S4.

Algorithm S4 Spherical K-median Spectral Clustering

Input An adjacency matrix $A_{n\times n}$, the number of communities K, and an approximating parameter $\delta > 0$ for K-median clustering.

Output A membership matrix $\hat{C}_{n\times K}$.

procedure SSC(A, K)

1. Calculate $\hat{U} \in \mathbb{R}^{n\times K}$ consisting of the leading K eigenvectors (ordered in absolute eigenvalue) of A.

2. Let $I_+ = \left\{i : \ \|\hat{U}_{i*}\| > 0\right\}$ and $\hat{U}^+ = (\hat{U}_{I_+*})$.

3. Let $\hat{U}'$ be the row-normalized version of $\hat{U}^+$.

4. Let $(\hat{C}^+, \hat{X}) \in \mathbb{C}_{n\times K} \times \mathbb{R}^{K\times K}$ such that $\|\hat{C}^+\hat{X} - \hat{U}'\|_{2,1} \leq (1+\delta) \min_{(C,X)\in\mathbb{C}_{n\times K}\times\mathbb{R}^{K\times K}} \|CX - \hat{U}'\|_{2,1}$.

5. Output $\hat{C}$ with $\hat{C}_{i*}$ being the corresponding row in $\hat{C}^+$ if $i \in I_+$, and $\hat{C}_{i*} = (1, 0, \ldots, 0)$ if $i \notin I_+$.

Theorem S2. *(Lei and Rinaldo, 2015) Let A be an adjacency matrix generated from a degree corrected blockmodel DCBM$(C_{n\times K}, P_{K\times K}, \psi_{n\times 1})$. As-*

sume that

1. $P = \alpha_n P_0$ *for some* $\alpha_n \geq \frac{1}{n}\log n$,

2. P_0 *has minimum absolute eigenvalue* $\geq \lambda > 0$,

3. $\max_{k,l} P_{0_{kl}} = 1$,

4. $\hat{C}$ *is the solution to spectral clustering using spherical* $(1+\delta)$*-approximate* K*-median (Algorithm S4) applied on* A *with* K *communities.*

Then there exists an absolute constant c *such that if*

$$(2.5+\delta)\frac{\sqrt{K}}{\lambda \tilde{n}_{min}\pi_{min}\sqrt{n\alpha_n}}\sqrt{\sum_{i=1}^{n}\tilde{\psi}_i^{-2}} < c, \tag{S1.5}$$

then, with probability at least $1-\frac{1}{n}$,

$$\delta(\hat{C}, C) = \frac{1}{n}\min_{E\in\mathbf{E}_K}\|\hat{C}E - C\|_0 \leq c^{-1}(2.5+\delta)\frac{\sqrt{K}}{\lambda\tilde{n}_{min}\sqrt{n\alpha_n}}\sqrt{\sum_{i=1}^{n}\tilde{\psi}_i^{-2}}. \tag{S1.6}$$

S2 Supporting results and proofs

S2.1 Proof of general bound

The following lemma from Mukherjee et al. (2021, Lemma S.6.3) bounds the tail probabilities of a hypergeometric random variable. It is used in the main proofs to obtain upper and lower bounds for the smallest and

largest community sizes in a random sample of nodes from the main network. Probability mass function of a Hypergeometric(n, d, N) is $f(x) = \frac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}}\mathbf{1}\left(x \in \{\max(0, n+d-N), \cdots, \min(n, d)\}\right)$.

Lemma S1 (Bound for Hypergeometric variable (Mukherjee et al., 2021))**.** *Let $D \sim$ Hypergeometric(n, d, N). Then for $\omega > 0$, we have*

$$\max\left\{P\left(D < (1-\omega)\frac{nd}{N}\right),\ P\left(D > (1+\omega)\frac{nd}{N}\right)\right\} \leq \exp\left(-\frac{\omega^2 nd}{2N(1+\frac{\omega}{3})}\right). \tag{S2.7}$$

Further if $0 < \omega < 1$,

$$\max\left\{P\left(D < (1-\omega)\frac{nd}{N}\right),\ P\left(D > (1+\omega)\frac{nd}{N}\right)\right\} \leq \exp\left(-\frac{\omega^2 nd}{4N}\right). \tag{S2.8}$$

Lemma S2. *Let C be any $n \times K$ matrix, and E be any $K \times K$ permutation matrix. Then*

$$\|C\|_0 = \sum_{i,j} |C_{ij}| = \|CE\|_0. \tag{S2.9}$$

Proof. Since multiplying with a permutation matrix changes only the positions of the elements inside a matrix, the lemma follows immediately. □

Lemma S3. *Let $\hat{o}_l$ be the size of the lth community in a random subgraph G_{S_0} of size o, induced by the nodes in $S_0 \subset S$ selected using SRSWOR from the n nodes in S. If π_l is the proportion of the lth community in the entire*

network G, $\pi_{min} = \min\{\pi_1, \ldots, \pi_K\}$, *and* $\pi_{max} = \max\{\pi_1, \ldots, \pi_K\}$, *then we have*

$$\hat{o}_{min} = \min_{l=1,\ldots,K} \hat{o}_l \geq \frac{o\pi_{min}}{1+\pi_{min}}, \text{ and} \tag{S2.10}$$

$$\hat{o}_{max} = \max_{l=1,\ldots,K} \hat{o}_l \leq \frac{o\pi_{max}}{1-\pi_{max}}, \tag{S2.11}$$

each with probability $\geq 1-\omega_o$, *where* $\omega_o = K\exp\left(-o\pi_{min}^3/(4(1+\pi_{min})^2)\right)$.

Proof. Since G_{S_0} is a random subgraph of size o from G, $\hat{o}_l \sim$ Hypergeometric (o, n_l, n), $l = 1, 2, \ldots, K$, where n_l is the size of the lth community in the entire network G. Note that $\pi_l = n_l/n$. Consider the following probability,

$$\begin{aligned}
P\left(\hat{o}_{min} < \frac{o\pi_{min}}{1+\pi_{min}}\right) &= P\left(\bigcup_{l=1}^{K}\left\{\hat{o}_l < \frac{o\pi_{min}}{1+\pi_{min}}\right\}\right) \\
&\leq \sum_{l=1}^{K} P\left(\hat{o}_l < \frac{o\pi_{min}}{1+\pi_{min}}\right) \\
&= \sum_{l=1}^{K} P\left(\hat{o}_l < o\pi_l\left\{1-\left(1-\frac{\pi_{min}}{\pi_l(1+\pi_{min})}\right)\right\}\right) \\
&\leq \sum_{l=1}^{K} \exp\left(-\left(1-\frac{\pi_{min}}{\pi_l(1+\pi_{min})}\right)^2\frac{o\pi_l}{4}\right) \\
&\qquad\qquad \left[\text{from (S2.8) as } 0 < \frac{\pi_{min}}{\pi_l(1+\pi_{min})} < 1\right] \\
&\leq K\exp\left(-\left(1-\frac{1}{1+\pi_{min}}\right)^2\frac{o\pi_{min}}{4}\right) \\
&= K\exp\left(-\frac{\pi_{min}^2}{(1+\pi_{min})^2}\frac{o\pi_{min}}{4}\right) \\
&= K\exp\left(-\frac{o\pi_{min}^3}{4(1+\pi_{min})^2}\right) = \omega_o.
\end{aligned} \tag{S2.12}$$

Now consider the following probability

$$
\begin{aligned}
P\left(\hat{o}_{max} > \frac{o\pi_{max}}{1-\pi_{max}}\right) &= P\left(\bigcup_{l=1}^{K}\left\{\hat{o}_l > \frac{o\pi_{max}}{1-\pi_{max}}\right\}\right) \\
&\leq \sum_{l=1}^{K} P\left(\hat{o}_l > \frac{o\pi_{max}}{1-\pi_{max}}\right) \\
&= \sum_{l=1}^{K} P\left(\hat{o}_l > o\pi_l\left\{1+\left(\frac{\pi_{max}}{\pi_l(1-\pi_{max})}-1\right)\right\}\right) \\
&\leq \sum_{l=1}^{K} \exp\left(-\left(\frac{\pi_{max}}{\pi_l(1-\pi_{max})}-1\right)^2 \frac{o\pi_l}{2\left(1+\frac{1}{3}\left(\frac{\pi_{max}}{\pi_l(1-\pi_{max})}-1\right)\right)}\right) \\
&\qquad\qquad \left[\text{from (S2.7) as } \frac{\pi_{max}}{\pi_l(1-\pi_{max})} > 1\right] \\
&\leq \sum_{l=1}^{K} \exp\left(-\left(\frac{1}{1-\pi_{max}}-1\right)^2 \frac{o\pi_l^2(1-\pi_{max})}{2\pi_{max}}\right) \\
&\leq \sum_{l=1}^{K} \exp\left(-\frac{o\pi_{max}\pi_l^2}{2(1-\pi_{max})}\right) \\
&\leq K\exp\left(-\frac{o\pi_{min}^3}{4(1+\pi_{min})^2}\right) = \omega_o. \qquad \text{(S2.13)}
\end{aligned}
$$

□

Proof of Theorem 1

Proof. Part 1: Note that $S_0 \subset S_q = S_0 \cup S_q'$, which implies

$$
\|\hat{C}_{S_{0*}}^{(q)} E_q^* - C_{S_{0*}}\|_0 \leq \|\hat{C}_{S_{q*}}^{(q)} E_q^* - C_{S_{q*}}\|_0 \leq \epsilon(o+m) \quad \text{w.p. } \geq 1-\alpha. \qquad \text{(S2.14)}
$$

Let $E_q' \in \mathbf{E}_K$ be any permutation matrix such that $E_q' \neq E_q^*$. Then

$$\begin{aligned}
\|\hat{C}^{(q)}_{S_{0*}} E_q' - C_{S_{0*}}\|_0 &= \|(\hat{C}^{(q)}_{S_{0*}} E_q' - \hat{C}^{(q)}_{S_{0*}} E_q^*) - (C_{S_{0*}} - \hat{C}^{(q)}_{S_{0*}} E_q^*)\|_0 \\
&\geq \|\hat{C}^{(q)}_{S_{0*}} E_q' - \hat{C}^{(q)}_{S_{0*}} E_q^*\|_0 - \|C_{S_{0*}} - \hat{C}^{(q)}_{S_{0*}} E_q^*\|_0 \\
&\geq 2\hat{o}_{min} - \|C_{S_{0*}} - \hat{C}^{(q)}_{S_{0*}} E_q^*\|_0 \\
&\quad \text{(since at least a pair of columns are different between } \hat{C}^{(q)}_{S_{0*}} E_q^* \text{ and } \hat{C}^{(q)}_{S_{0*}} E_q') \\
&\geq \frac{2o\pi_{min}}{1+\pi_{min}} - \epsilon(o+m) \ \text{ w.p. } \geq 1-\omega_o-\alpha \text{ (from Lemma S3 and Condition (3.3))} \\
&\geq \epsilon(o+m) \ \text{ w.p. } \geq 1-\omega_o-\alpha. \text{ (from Condition (3.4))}
\end{aligned} \tag{S2.15}$$

Thus, combining (S2.14) and (S2.15), we have

$$\begin{aligned}
&\|\hat{C}^{(q)}_{S_{0*}} E_q' - C_{S_{0*}}\|_0 \geq \epsilon(o+m) \geq \|\hat{C}^{(q)}_{S_{0*}} E_q^* - C_{S_{0*}}\|_0 \ \text{ w.p. } \geq 1-\omega_o-2\alpha \quad \text{for each} \quad E_q' \neq E_q^* \\
\implies & E_q^* = \underset{E\in\mathbf{E}_K}{\arg\min} \|\hat{C}^{(q)}_{S_{0*}} E - C_{S_{0*}}\|_0 \ \text{ w.p. } \geq 1-\omega_o-2\alpha \quad \text{for each} \quad E_q' \neq E_q^*.
\end{aligned} \tag{S2.16}$$

Part 2: Consider the following quantity

$$\begin{aligned}
\|\hat{C}^{(q)}_{S_{0*}} E_q^* E_1^{*-1} - \hat{C}^{(1)}_{S_{0*}}\|_0 &= \|(\hat{C}^{(q)}_{S_{0*}} E_q^* E_1^{*-1} - C_{S_{0*}} E_1^{*-1}) + (C_{S_{0*}} E_1^{*-1} - \hat{C}^{(1)}_{S_{0*}})\|_0 \\
&\leq \|\hat{C}^{(q)}_{S_{0*}} E_q^* E_1^{*-1} - C_{S_{0*}} E_1^{*-1}\|_0 + \|C_{S_{0*}} E_1^{*-1} - \hat{C}^{(1)}_{S_{0*}}\|_0 \\
&= \|\hat{C}^{(q)}_{S_{0*}} E_q^* - C_{S_{0*}}\|_0 + \|\hat{C}^{(1)}_{S_{0*}} E_1^* - C_{S_{0*}}\|_0 \text{ (from Lemma S2 as } E_1^* \in \mathbf{E}_K) \\
&\leq 2\epsilon(o+m) \ \text{ w.p. } \geq 1-2\alpha.
\end{aligned} \tag{S2.17}$$

Also, for any $E' \in \mathbf{E}_K$ such that $E' \neq E_q^* E_1^{*-1}$, we have

$$\begin{aligned}
\|\hat{C}^{(q)}_{S_0*} E' - \hat{C}^{(1)}_{S_0*}\|_0 &= \|(\hat{C}^{(q)}_{S_0*} E' - \hat{C}^{(q)}_{S_0*} E_q^* E_1^{*-1}) - (\hat{C}^{(1)}_{S_0*} - \hat{C}^{(q)}_{S_0*} E_q^* E_1^{*-1})\|_0 \\
&\geq \|\hat{C}^{(q)}_{S_0*} E' - \hat{C}^{(q)}_{S_0*} E_q^* E_1^{*-1}\|_0 - \|\hat{C}^{(1)}_{S_0*} - \hat{C}^{(q)}_{S_0*} E_q^* E_1^{*-1}\|_0 \\
&\geq 2\hat{o}_{min} - 2\epsilon(o+m) \quad \text{w.p.} \quad \geq 1 - 2\alpha \\
&\geq \frac{2o\pi_{min}}{1+\pi_{min}} - 2\epsilon(o+m) \quad \text{w.p.} \quad \geq 1 - \omega_o - 2\alpha \\
&\geq 2\epsilon(o+m) \quad \text{w.p.} \quad \geq 1 - \omega_o - 2\alpha. \qquad \text{(S2.18)}
\end{aligned}$$

Thus, combining (S2.17) and (S2.18), we have

$$\begin{aligned}
&\|\hat{C}^{(q)}_{S_0*} E' - \hat{C}^{(1)}_{S_0*}\|_0 \geq 2\epsilon(o+m) \geq \|\hat{C}^{(q)}_{S_0*} E_q^* E_1^{*-1} - \hat{C}^{(1)}_{S_0*}\|_0 \\
&\qquad\qquad \text{w.p.} \quad \geq 1 - \omega_o - 4\alpha \quad \text{for each} \quad E' \neq E_q^* E_1^{*-1} \\
&\implies E_q^* E_1^{*-1} = \underset{E \in \mathbf{E}_K}{\arg\min} \|\hat{C}^{(q)}_{S_0*} E - \hat{C}^{(1)}_{S_0*}\|_0 \quad \text{w.p.} \quad \geq 1 - \omega_o - 4\alpha \quad \text{for each} \quad E' \neq E_q^* E_1^{*-1}. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(S2.19)}
\end{aligned}$$

□

Proof of Theorem 2

Proof. Consider the following quantity

$$
\begin{aligned}
\mathcal{M}(\hat{C}, C) &= \min_{E \in \mathbf{E}_K} \|\hat{C}E - C\|_0 \\
&\leq \|\hat{C}E_1^* - C\|_0 \\
&= \|\hat{C}^{(1)}_{S_{1*}} E_1^* - C_{S_{1*}}\|_0 + \|\hat{C}^{(2)}_{S'_{2*}} E_2 E_1^* - C_{S'_{2*}}\|_0 + \cdots + \|\hat{C}^{(s)}_{S'_{s*}} E_s E_1^* - C_{S'_{s*}}\|_0 \\
&\leq \|\hat{C}^{(1)}_{S_{1*}} E_1^* - C_{S_{1*}}\|_0 + \|\hat{C}^{(2)}_{S'_{2*}} E_2^* E_1^{*-1} E_1^* - C_{S'_{2*}}\|_0 + \cdots + \|\hat{C}^{(s)}_{S'_{s*}} E_s^* E_1^{*-1} E_1^* - C_{S'_{s*}}\|_0 \\
&\qquad \text{w.p. } \geq 1 - (s-1)(\omega_o + 4\alpha) \text{ (from (3.6))} \\
&= \|\hat{C}^{(1)}_{S_{1*}} E_1^* - C_{S_{1*}}\|_0 + \|\hat{C}^{(2)}_{S'_{2*}} E_2^* - C_{S'_{2*}}\|_0 + \cdots + \|\hat{C}^{(s)}_{S'_{s*}} E_s^* - C_{S'_{s*}}\|_0 \\
&\qquad \text{w.p. } \geq 1 - (s-1)(\omega_o + 4\alpha) \\
&\leq \|\hat{C}^{(1)}_{S_{1*}} E_1^* - C_{S_{1*}}\|_0 + \|\hat{C}^{(2)}_{S_{2*}} E_2^* - C_{S_{2*}}\|_0 + \cdots + \|\hat{C}^{(s)}_{S_{s*}} E_s^* - C_{S_{s*}}\|_0 \\
&\qquad \text{w.p. } \geq 1 - (s-1)(\omega_o + 4\alpha) \\
&= \mathcal{M}(\hat{C}^{(1)}_{S_{1*}}, C_{S_{1*}}) + \mathcal{M}(\hat{C}^{(2)}_{S_{2*}}, C_{S_{2*}}) + \cdots + \mathcal{M}(\hat{C}^{(s)}_{S_{s*}}, C_{S_{s*}}) \\
&\qquad \text{w.p. } \geq 1 - (s-1)(\omega_o + 4\alpha) \\
&\leq s\epsilon(o+m) \text{ from Condition (3.3)} \\
&\qquad \text{w.p. } \geq 1 - (s-1)(\omega_o + 4\alpha) - s\alpha.
\end{aligned}
\tag{S2.20}
$$

Hence the result follows. □

S2.2 Detailed Results and Proofs on Data Usage Proportion

We present a result on the expected proportion of node pairs used in `SONNET` and conduct simulations to establish the effect of repetition in increasing the data usage. For the results in this section, we assume that the network adjacency matrix is not symmetric, i.e., the node pairs (i, j) and (j, i) are different, and self-loops are allowed, i.e., the node pairs (i, i) may be non-zero.

It is easier to obtain the proportion of unused node pairs in `SONNET` with s subgraphs, overlapping size o and r repetitions. Borrowing notations from Algorithm 1, the set of node pairs not used in the base step ($r = 0$) of `SONNET` is $W_0 \coloneqq \cup_{1\le p\neq q\le s}(S'_p \times S'_q)$, and similarly, the set of node pairs not used in the ρth repetition step of `SONNET` is $W_\rho \coloneqq \cup_{1\le p\neq q\le s}(S'_{p\rho} \times S'_{q\rho})$. By the structure of the algorithms, all node pairs from $S_0 \times S$ is used in `SimpleSONNET` and the 0th repetition step of `SONNET`. A node pair (i, j) is defined as unused for the first ρ repetition steps of `SONNET` if that node pair is not included in `SONNET` up to that point, i.e., $(i, j) \in \cap_{\mu=0}^{\rho} W_\mu$. Lemma S4 presents a result on the expected number of unused node pairs for the first ρth repetition steps. Theorem 3 follows directly from this lemma.

Lemma S4. *Let `SONNET`, with a suitable community detection algorithm and parameters s, o, and r, be applied to a network with nodes in S. Let*

$X_\rho = |\cap_{\mu=0}^{\rho} W_\mu|$ *be the number of unused node pairs up to the* ρ*th repetition. Then the expectation of* X *is given by*

$$E\left[X_\rho\right] = (n-o)^2 \left(\frac{s-1}{s}\right)^{\rho+1}. \tag{S2.21}$$

Proof. For any $0 \leq \rho \leq r$ and node pair $(i,j) \in (S \setminus S_0) \times (S \setminus S_0)$, consider the following probability:

$$\begin{aligned}
&P\left[(i,j) \text{ is unused up to the } \rho\text{th repetition}\right] \\
=&P\left[(i,j) \in \bigcap_{\mu=0}^{\rho} W_\mu\right] \\
=&\prod_{\mu=0}^{\rho} P\left[(i,j) \in W_\mu\right] \quad \begin{array}{l}\text{(as the non-overlapping parts are randomly} \\ \text{shuffled at each repetition step)}\end{array} \\
=&\prod_{\mu=0}^{\rho} P\left[(i,j) \in \bigcup_{1 \leq p \neq q \leq s} (S'_{p\mu} \times S'_{q\mu})\right] \\
=&\prod_{\mu=0}^{\rho} \sum_{p=1}^{s} P\left[j \in S'_{p\mu}, i \notin S'_{p\mu}\right] \\
=&\prod_{\mu=0}^{\rho} \sum_{p=1}^{s} P\left[j \in S'_{p\mu}\right] P\left[i \notin S'_{p\mu}\right] \quad \begin{array}{l}\text{(due to random assignments} \\ \text{of nodes into subgraphs)}\end{array} \\
=&\prod_{\mu=0}^{\rho} \sum_{p=1}^{s} \frac{1}{s} \times \frac{s-1}{s} \\
=&\left(\frac{s-1}{s}\right)^{\rho+1}.
\end{aligned} \tag{S2.22}$$

Thus, from (S2.22),

$$E[X_\rho] = \sum_{(i,j)\in(S\setminus S_0)\times(S\setminus S_0)} P\left[(i,j) \in \bigcap_{\mu=0}^{\rho} W_\mu\right] = (n-o)^2 \left(\frac{s-1}{s}\right)^{\rho+1}. \quad \text{(S2.23)}$$

□

The expected data usage proportion is also a function of the number of subgraphs s, and it is smaller for larger s given a fixed value of ρ. We simulated the divide and the repetition steps of `SONNET` to estimate the proportion of used node pairs in a 10000-node network. Overlapping size o was taken as 1000, and we obtained line charts of estimated and theoretical proportions of used node pairs for four different values of s: 10, 20, 50, and 100. The number of repetitions r ranges from 0 (`SimpleSONNET`) to 50. The plot is presented in Figure 1.

From Figure 1, we observe that for all the cases, the expected and simulated data usage proportions are visibly indistinguishable. For smaller values of s, a higher data usage proportion is achieved with fewer repetitions r. For larger values of s, large values of r are required to achieve a reasonable data usage proportion.

Regarding the data usage maximization with time constraint for parameter selection as in Section 3.3, one may need to provide a search space for

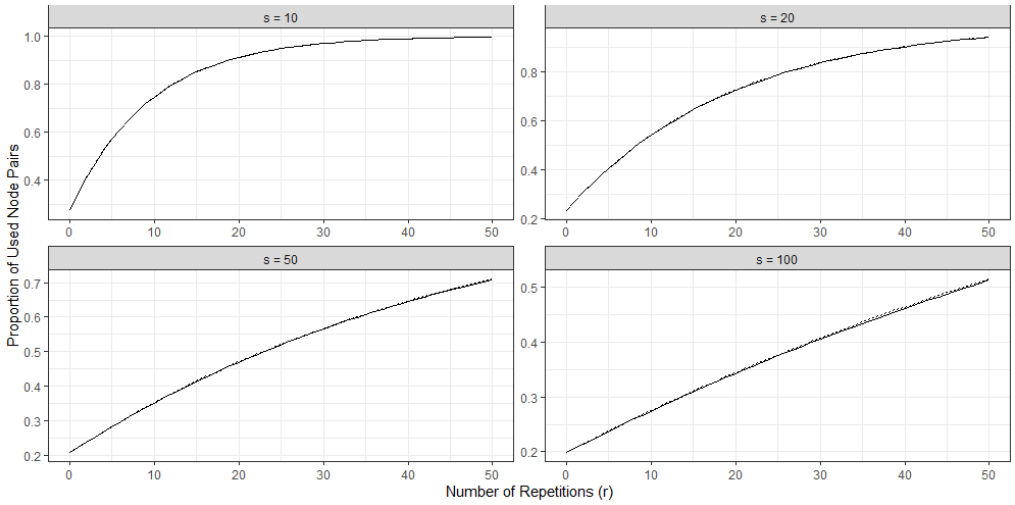

Figure 1: Estimated (solid line) and expected (dashed line) data usage proportions over the number of repetitions (r): Simulations are based on $n = 10000$ nodes and overlapping size $o = 1000$

s and o depending on the optimizer. The search space can be decided based on the computing resources available. In the case that `SONNET` is parallelized over *ncore* processors, barring a small overhead due to parallelization, computation complexity is approximately divided by *ncore*. Thus, one may replace q by $q \times ncore$ in the previous expression.

S2.3 Proof of bound for `SONNET` with spectral clustering on SBM

Proof of Theorem 4

Proof. Note that we use the superscript $(\mathbb{S}_d)$ on any quantity to indicate the feature of the subgraph $G_{\mathbb{S}_d}$. Also note that the quantities λ and α_n remain

the same for the random subgraph $G_{\mathbb{S}_d}$. Conditions 1, 2, and 3 of Theorem S1 are also satisfied by $G_{\mathbb{S}_d}$ as they are satisfied by G. The subgraph $G_{\mathbb{S}_d}$ can be treated as a network modeled by SBM$(C^{(\mathbb{S}_d)}_{d\times K}, P_{K\times K})$.

From a direct application of Lemma S3, we have

$$(2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} \leq (2+\delta)\frac{K(1+\pi_{min})^2}{\lambda^2\pi^2_{min}d\alpha_n} \quad \text{w.p.} \ \geq 1-\omega_d,$$

where $\omega_d = K\exp\left(-\frac{d\pi^3_{min}}{4(1+\pi_{min})^2}\right)$. Thus,

$$\begin{aligned}&(2+\delta)\frac{K(1+\pi_{min})^2}{\lambda^2\pi^2_{min}d\alpha_n} < c\\ \Longrightarrow\ &(2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} < c \quad \text{w.p.} \ \geq 1-\omega_d.\end{aligned}$$

Consider the following probability

$$\begin{aligned}&\boldsymbol{P}\left(\delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d*}) > c^{-1}(2+\delta)\frac{K\pi^{(\mathbb{S}_d)}_{max}}{\pi^{(\mathbb{S}_d)2}_{min}\lambda^2 d\alpha_n}\right)\\ =&\boldsymbol{P}\left(\delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d*}) > c^{-1}(2+\delta)\frac{K\pi^{(\mathbb{S}_d)}_{max}}{\pi^{(\mathbb{S}_d)2}_{min}\lambda^2 d\alpha_n}\middle|(2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} < c\right)\boldsymbol{P}\left((2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} < c\right)\\ &+\boldsymbol{P}\left(\delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d*}) > c^{-1}(2+\delta)\frac{K\pi^{(\mathbb{S}_d)}_{max}}{\pi^{(\mathbb{S}_d)2}_{min}\lambda^2 d\alpha_n}\middle|(2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} \geq c\right)\boldsymbol{P}\left((2+\delta)\frac{K}{\lambda^2\pi^{(\mathbb{S}_d)2}_{min}d\alpha_n} \geq c\right)\\ \leq&\frac{1}{d}\times 1 + 1\times\omega_d\end{aligned}$$

(the first bound is from Theorem S1, and second bound is from the previous equation).

Therefore, $\delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d*}) \leq c^{-1}(2+\delta)\frac{K\pi^{(\mathbb{S}_d)}_{max}}{\pi^{(\mathbb{S}_d)2}_{min}\lambda^2 d\alpha_n}$ w.p. $\geq 1-\frac{1}{d}-\omega_d$. Combining with Lemma S3, we have

$$\delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d*}) \leq c^{-1}(2+\delta)\frac{K\pi_{max}(1+\pi_{min})^2}{\pi^2_{min}(1-\pi_{max})\lambda^2 d\alpha_n} \quad \text{w.p.} \ \geq 1-\frac{1}{d}-3\omega_d.$$

□

Proof of Theorem 5

Proof. Condition (3.13) ensures that Theorem 4 can be applied. Replacing $d = o + m$ in Theorem 4, we can see that Condition (3.3) in Theorem 1 is satisfied with $\alpha = \frac{1}{o+m} + 3\omega_{o+m}$, and $\epsilon = c^{-1}(2+\delta)\frac{K\pi_{max}(1+\pi_{min})^2}{\pi_{min}^2(1-\pi_{max})\lambda^2(o+m)\alpha_n}$. Condition (3.14) translates to Condition (3.4) in Theorem 1. Thus, combining Theorem 4 and Theorem 2, we have the final bound in Theorem 5. □

S2.4 Proof of bound for SONNET with spherical K-median spectral clustering on DCBM

Here, we present a result on a bound on the sum from an SRSWOR sample, given by Serfling (1974). Consider a population P containing N elements $P = \{p_1, \ldots, p_N\}$, with $p_i \in \mathbb{R}$. Let a and b be the minimum and the maximum value in P, respectively, and $\mu = \frac{1}{N}\sum_{i=1}^{N} p_i$ be the population mean. Let $1 \leq i \leq n \leq N$, and X_i be the ith SRSWOR draw from the population P. Define $M_n = \sum_{i=1}^{n} X_i$. Then the following lemma holds.

Lemma S5. *(Finite sampling bound (Serfling, 1974)) For* $1 \leq n \leq N$, $\lambda > 0$, *and* M_n *being the sum in SRSWOR,*

$$\max\left\{\boldsymbol{P}\left(\sqrt{n}\left(\frac{M_n}{n}-\mu\right)\geq\lambda\right),\ \boldsymbol{P}\left(\sqrt{n}\left(\frac{M_n}{n}-\mu\right)\leq-\lambda\right)\right\}\leq\exp\left(\frac{-2\lambda^2}{(1-f_n)(b-a)^2}\right), \tag{S2.24}$$

where $f_n = \frac{n-1}{N}$.

Lemma S6. *Suppose* $G_{\mathbb{S}_d}$ *is a random subgraph of size* d *from a network* G *of size* n *that is generated by DCBM. Assume* $\tilde{n}_{min}^{(\mathbb{S}_d)}$ *is defined similarly for* $G_{\mathbb{S}_d}$ *as* $\tilde{n}_{min}$ *for* G. *Then with probability* $\geq 1 - \frac{K}{n}$,

$$\tilde{n}_{min}^{(\mathbb{S}_d)} \geq \frac{d}{n}\tilde{n}_{min} - \gamma^*, \tag{S2.25}$$

where $\gamma^* = \left(\frac{d(n-d+1)\log n}{2n}\right)^{\frac{1}{2}}$.

Proof. Note that for $1 \leq l \leq K$, $\tilde{n}_l^{(\mathbb{S}_d)} = \sum_{i\in\mathbb{S}_d} \psi_i^2 \mathbf{1}\{i \in G_l\}$ is a sum of an SRSWOR sample of size d from the population $P = \{\psi_1^2\mathbf{1}\{1 \in G_l\}, \psi_2^2\mathbf{1}\{2 \in G_l\}, \ldots, \psi_n^2\mathbf{1}\{n \in G_l\}\}$. Since $\max_{1\leq i\leq n} \psi_i = 1$, the maximum and the minimum values in the population are respectively 1 and 0. Define $\tilde{\mu}_l = \frac{1}{n}\sum_{i=1}^{n} \psi_i^2 \mathbf{1}\{i \in G_l\}$.

For any $\gamma > 0$, using union bound and Lemma S5,

$$\begin{aligned}
\boldsymbol{P}\left(\tilde{n}_{min}^{(\mathbb{S}_d)} < \frac{d}{n}\tilde{n}_{min} - \gamma\right) &= \boldsymbol{P}\left(\bigcup_{l=1}^{K}\left\{\tilde{n}_l^{(\mathbb{S}_d)} < \frac{d}{n}\tilde{n}_{min} - \gamma\right\}\right) \\
&\leq \sum_{l=1}^{K} \boldsymbol{P}\left(\tilde{n}_l^{(\mathbb{S}_d)} < \frac{d}{n}\tilde{n}_{min} - \gamma\right) \\
&= \sum_{l=1}^{K} \boldsymbol{P}\left(\sqrt{d}\left(\frac{\tilde{n}_l^{(\mathbb{S}_d)}}{d} - \tilde{\mu}_l\right) < \sqrt{d}\left(\frac{\tilde{n}_{min}}{n} - \tilde{\mu}_l\right) - \frac{\gamma}{\sqrt{d}}\right) \\
&\left[\text{from Lemma S5 as } \frac{\tilde{n}_{min}}{n} - \tilde{\mu}_l = \min_{1<l<K} \frac{1}{n}\sum_{i=1}^{n}\psi_i^2\mathbf{1}\{i \in G_l\} - \frac{1}{n}\sum_{i=1}^{n}\psi_i^2\mathbf{1}\{i \in G_l\} \leq 0,\ \forall\ l\right] \qquad \text{(S2.26)} \\
&\leq \sum_{l=1}^{K} \exp\left(\frac{-2\left\{\sqrt{d}\left(\frac{\tilde{n}_{min}}{n} - \tilde{\mu}_l\right) - \frac{\gamma}{\sqrt{d}}\right\}^2}{1 - f_d}\right) \\
&\leq K\exp\left(\frac{-2\gamma^2}{d(1-f_d)}\right) \\
&= K\exp\left(\frac{-2n\gamma^2}{d(n-d+1)}\right). \qquad \text{(S2.27)}
\end{aligned}$$

Set $\gamma = \left(\frac{d(n-d+1)\log n}{2n}\right)^{\frac{1}{2}}$ and the result follows. □

Lemma S7. *Suppose $G_{\mathbb{S}_d}$ is a random subgraph of size d from a network G of size n that is generated by DCBM. Define $S_{\tilde{\psi}} = \sum_{i=1}^{n}\left(\psi_i^{-2}\sum_{l\in G_{g_i}}\psi_l^2\right)$, and $R_{\tilde{\psi}} = \max_{i=1,\ldots,n}\psi_i^{-2}\sum_{l\in G_{g_i}}\psi_l^2 - \min_{i=1,\ldots,n}\psi_i^{-2}\sum_{l\in G_{g_i}}\psi_l^2$. Then with probability $\geq 1 - \frac{1}{n}$,*

$$\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2} \leq \frac{d}{n}S_{\tilde{\psi}} + \eta^*, \qquad \text{(S2.28)}$$

where $\eta^* = \left(\frac{d(n-d+1)\log n}{2n}\right)^{\frac{1}{2}} R_{\tilde{\psi}}$.

Proof. For any $\eta > 0$, using Lemma S5, we have

$$\boldsymbol{P}\left(\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2} > \frac{d}{n}S_{\tilde{\psi}} + \eta\right) = \boldsymbol{P}\left(\sqrt{d}\left(\frac{1}{d}\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2} - \frac{1}{n}S_{\tilde{\psi}}\right) > \frac{\eta}{\sqrt{d}}\right)$$
$$\leq \exp\left(\frac{-2\eta^2}{d(1-f_d)R_{\tilde{\psi}}^2}\right). \tag{S2.29}$$

Set $\eta = \left(\frac{d(n-d+1)\log n}{2n}\right)^{\frac{1}{2}} R_{\tilde{\psi}}$ and the result follows. □

Proof of Theorem 6

Proof. Define the events $A = \left\{\tilde{n}_{min}^{(\mathbb{S}_d)} \geq \frac{d}{n}\tilde{n}_{min} - \gamma^*\right\}$, $B = \left\{\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2} \leq \frac{d}{n}S_{\tilde{\psi}} + \eta^*\right\}$, and $C = \left\{\pi_{min}^{(\mathbb{S}_d)} \geq \frac{\pi_{min}}{1+\pi_{min}}\right\}$.

Note that $\boldsymbol{P}(A) \geq 1 - \frac{K}{n}$ from Lemma S6, $\boldsymbol{P}(B) \geq 1 - \frac{1}{n}$ from Lemma S7, and $P(C) \geq 1 - \omega_d$ from Lemma S3. Also, $A \cap B \cap C$ implies the event

$$D = \left\{(2.5+\delta)\frac{\sqrt{K}}{\lambda\tilde{n}_{min}^{(\mathbb{S}_d)}\pi_{min}^{(\mathbb{S}_d)}\sqrt{n\alpha_n}}\sqrt{\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2}} \leq (2.5+\delta)\frac{\sqrt{K\left(\frac{d}{n}S_{\tilde{\psi}} + \eta^*\right)}(1+\pi_{min})}{\lambda\left(\frac{d}{n}\tilde{n}_{min} - \gamma^*\right)\pi_{min}\sqrt{n\alpha_n}}\right\}.$$

Then

$$\begin{aligned}\boldsymbol{P}(D) &\geq P(A \cap B \cap C))\\ &= 1 - \boldsymbol{P}(A^c \cup B^c \cup C^c)\\ &\geq 1 - \boldsymbol{P}(A^c) - \boldsymbol{P}(B^c) - \boldsymbol{P}(C^c)\\ &\geq 1 - \frac{K}{n} - \frac{1}{n} - \omega_d. \end{aligned} \tag{S2.30}$$

Define the event $E = \left\{(2.5+\delta)\frac{\sqrt{K}}{\lambda\tilde{n}_{min}^{(\mathbb{S}_d)}\pi_{min}^{(\mathbb{S}_d)}\sqrt{n\alpha_n}}\sqrt{\sum_{i\in\mathbb{S}_d}\tilde{\psi}_i^{-2}} < c\right\}$. Then

from Condition (3.16), we have

$$P(E) \geq 1 - \frac{K+1}{n} - \omega_d. \tag{S2.31}$$

Now, define the event $F = \left\{ \delta(\hat{C}^{(\mathbb{S}_d)}, C_{\mathbb{S}_d *}) \leq c^{-1}(2.5+\delta) \frac{\sqrt{K}}{\lambda \tilde{n}_{min}^{(\mathbb{S}_d)} \sqrt{n\alpha_n}} \sqrt{\sum_{i \in \mathbb{S}_d} \tilde{\psi}_i^{-2}} \right\}$. Then

$$\begin{aligned} \boldsymbol{P}(F^c) &= \boldsymbol{P}(F^c|E)\boldsymbol{P}(E) + \boldsymbol{P}(F^c|E^c)P(E^c) \\ &\leq \frac{1}{d} \times 1 + 1 \times \left(\frac{K+1}{n} + \omega_d \right), \end{aligned} \tag{S2.32}$$

where the bound for the first term on the right hand side of (S2.32) follows from Theorem S2, and the second bound is from (S2.31). Note that $A \cap B \cap F$ implies the final bound in the theorem. Thus, the bound holds w.p. $\geq 1 - \frac{1}{d} - \omega_d - \frac{2(K+1)}{n}$. □

Proof of Theorem 7

Proof. Condition (3.18) ensures that Theorem 6 can be applied. Replacing $d = o + m$ in Theorem 6, we can see that Condition (3.3) in Theorem 1 is satisfied with $\alpha = \frac{1}{o+m} + \omega_{o+m} + \frac{2(K+1)}{n}$, and $\epsilon = c^{-1}(2.5 + \delta) \frac{\sqrt{K(\frac{o+m}{n} S_{\tilde{\psi}} + \eta^*)}}{\lambda\left(\frac{o+m}{n} \tilde{n}_{min} - \gamma^*\right)\sqrt{(o+m)\alpha_n}}$. Condition (3.19) translates to Condition (3.4) in Theorem 1. Thus, combining Theorem 6 and Theorem 2, we have the final bound in Theorem 7. □

S2.5 Comparison with Other Divide and Conquer Methods

We compare the performance of SONNET with a divide and conquer algorithm proposed by Mukherjee et al. (2021) in this section. Mukherjee et al. (2021) suggested two divide and conquer methods: PACE and GALE, with results on their consistency. GALE applies the parent algorithm $\mathcal{A}$ on T random subgraphs, each of size p, and matches the output labels along a traversal on the hypergraph of the subgraphs, where two subgraphs are connected by an edge if they have at least $p_1 = p^2/2n$ common nodes. GALE only uses the outputs from subgraphs that have at least τ common nodes with the union of the previous subgraphs along the traversal. SONNET and GALE are similar in the sense that they both apply the parent algorithm on multiple subgraphs of the network. However, GALE does not fix the overlap part and depends on the subgraph traversal for stitching. On the contrary, SONNET fixes the overlap part and partitions the remaining network to ensure that label matching can be done between the subgraphs without the need for traversing and searching for subgraphs with large randomly occurring overlaps. This gives SONNET more control for the parameter tuning step to achieve a lower error rate in less time compared to GALE. We restrict the comparison to GALE since PACE is structurally different from SONNET, as it stitches the output clustering matrices from each subgraph instead of the

community labels and requires an additional K-means clustering in the end to obtain the community labels. Also, in the numerical examples considered in this paper, `PACE` achieved higher error rates in longer times compared to `GALE`.

Here, we restate the theorem on the general error bound of `GALE` with T subgraphs each of size p and the threshold parameter $\tau = \theta Tp/n$, for a suitable $0 < \theta < 1$.

Theorem S3. *(Mukherjee et al., 2021, Theorem 3.2) Let* $0 < \theta < 1$ *and* $b, b' > 0$. *Let* $\hat{C}^{GALE}$ *be the output community labels of* ***GALE*** *applied with* T *subgraphs each of size* p *and threshold parameter* $\tau = \theta Tp/n$ *on a network of size* n *with a parent algorithm* $\mathcal{A}$ *that labels any random* p *subgraph with error* $\leq p^2\pi_{min}/24n$ *with probability at least* $1 - \delta$. *Let* $p \geq C\sqrt{\frac{n \log n}{\pi_{min}}}$, $T \geq C' n \log n/p$, *where* C *and* C' *are absolute constants that depend on* r, r', *and* θ. *Then, with probability at least* $1 - T\delta - O(n^{-r'})$, *the worst case error rate over all possible traversals of spanning trees of the hypergraph of the subgraphs is*

$$\delta(\hat{C}^{GALE}, C) \leq \frac{1}{\theta T} \sum_{l=1}^{T} \delta(\hat{C}^{(l)}, C^{(l)}) + O(n^{-r}), \tag{S2.33}$$

where $\hat{C}^{(l)}$ *and* $C^{(l)}$ *are respectively the estimated and the true community labels for the* l*th subgraph of* ***GALE****.*

To make the error bound of `GALE` (S2.33) comparable to the error bound of `SimpleSONNET` (3.7), we need to assume the same parent algorithm for both methods that satisfies Assumptions (3.3), (3.4), and $\epsilon < p^2\pi_{min}/24n$. Then, the two bounds reduce to

$$\texttt{SimpleSONNET}\text{: } \delta(\hat{C}^{SS}, C) \leq \frac{s(o+m)}{n}\epsilon \text{ with high probability,} \quad \text{(S2.34)}$$

$$\texttt{GALE}\text{: } \delta(\hat{C}^{GALE}, C) \leq \frac{1}{\theta}\epsilon = \frac{Tp}{\tau n}\epsilon \text{ with high probability.} \quad \text{(S2.35)}$$

Note that both error bounds share a similar structure. In `SimpleSONNET`, $\mathcal{A}$ is applied on s subgraphs each of size $(o+m)$ and in `GALE`, $\mathcal{A}$ is applied on T subgraphs each of size p. Thus, $s(o+m)/n$ and Tp/n can be interpreted as the total number of nodes used in `SimpleSONNET` and `GALE`, respectively, compared to the number of nodes in the entire network. Usually, these quantities need to be greater than 1 to produce accurate results for both methods. However, `GALE` has an additional threshold parameter τ to determine if any subgraphs have enough overlap with the union of the previous subgraphs along the traversal to contain sufficient information for sequential label matching. Although a large value of τ may improve the error bound substantially, it might lead to the rejection of many subgraphs and their traversals requiring large values of T and p that may slow down `GALE`. Even when τ is small, `GALE` usually needs larger subgraphs to

ensure that the random overlap between them is substantial. Fixing the overlapping part in `SimpleSONNET` helps it get by with smaller subgraphs and eliminates the need for searching a suitable traversal of the subgraphs, saving computation time. Also, for any selection of parameters T, p, τ of `GALE`, one can obtain s, o of `SimpleSONNET` such that the bound in (S2.34) is tighter than (S2.35).

S2.6 Error rates and runtimes for different values of `SONNET` parameters

Here, we present plots of error rates and computation times of `SONNET` against its parameters s, o, and r, keeping the other two fixed. The simulations are for the 10000-node SBM setup as in Section 3.4. We considered five different values of s: 10, 15, 20, 30, and 50. The values of o were 10, 500, and 1000 and the values of r were 0,2, and 5. We computed error rates and computation times for `SONNET` with all possible combinations of these values of s, o, and r. Figure 2 contains the plot. The plots on the left side of Figure 2 contains error rates against the parameter and the plots on the right side contain the computation times. The top levels of plots are against the number of subgraphs s, the medium level against the overlap size o, and the bottom level against the number of repetitions r.

Inside each plot, each line represents the error rates and computation times against their corresponding parameter , keeping the other two parameters fixed.

From Figure 2, it can be seen that when the overlap size and the number of repetitions are kept fixed, the error rate increases with s, but the computation time initially decreases and then increases. Error rates are observed to decrease as o and r increase, when the other two parameters are kept fixed. Computation times are observed to increase as o and r increase. These patterns indicate that one may choose a moderate value of s, where the error rate is low and the computation time has not increased by much. For both the overlapping part and the number of repetitions, it is usually the higher the better.

S2.7 Details on Implementation of `SONNET` and Real Data Examples

Implementation Details All the computations were performed on R version 4.0.3 on a university campus cluster equipped with Intel Xeon X5355 CPU with operating frequency 2.66 GHz. We ran all the codes with 20 such processors and 12 GB of memory per processor. The R codes for generating networks from SBM and DCBM, different variants of spectral clustering,

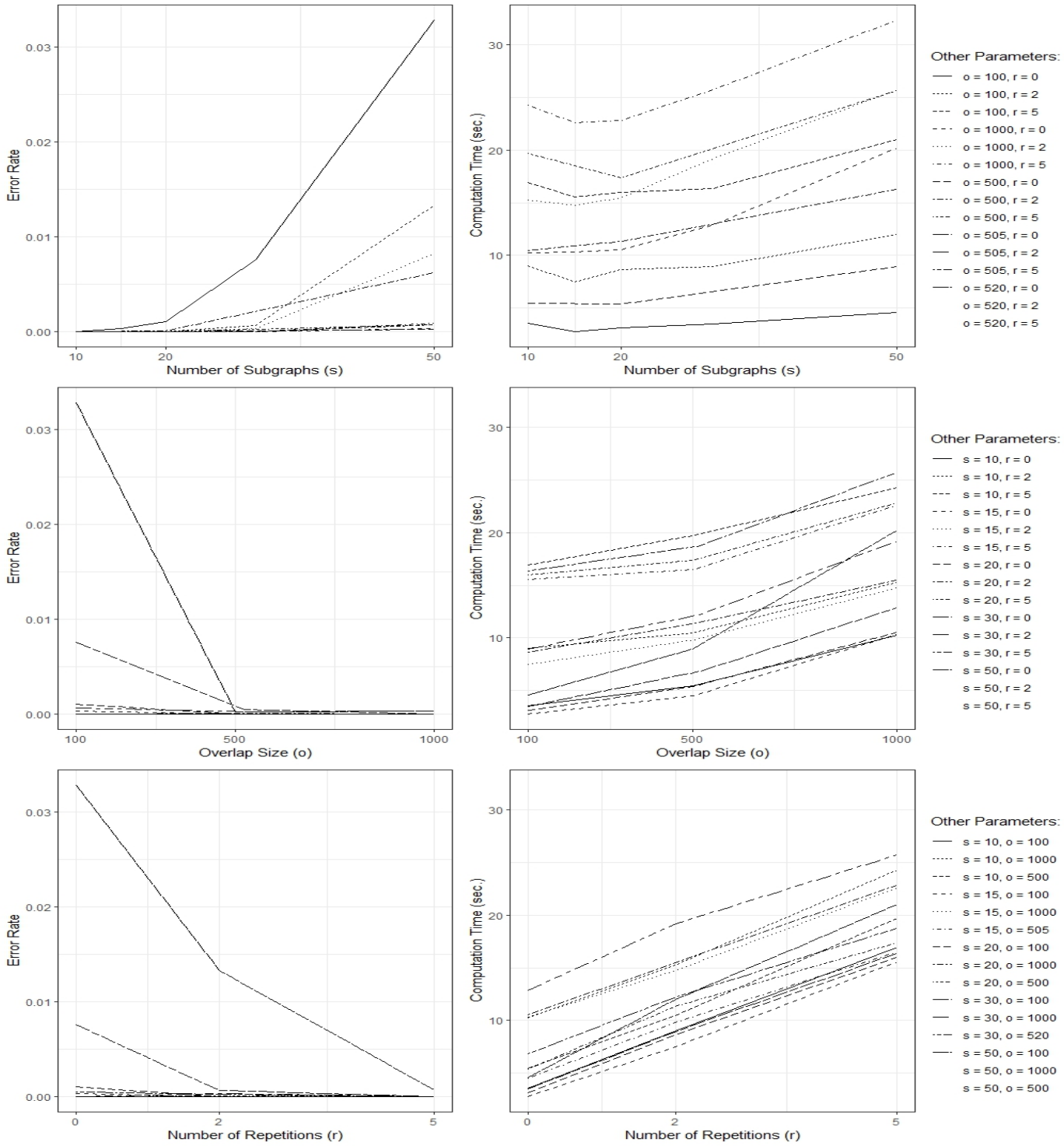

Figure 2: Error rates and computation times of `SONNET` against the parameters s, o, and r, individually, with the other two kept fixed. All the error rates and computation times are averages from 100 simulation of 10000-node SBM network with $K = 5$ communities, $p^{(intra)} = 0.2$, and $p^{(inter)} = 0.05$.

label matching, `SONNET`, and `GALE` were written by us. Eigen decomposition was done using the base 'eigen' function, K-means clustering used 'kmeans' function, and spherical K-median clustering used 'pam' function from the package 'pam'. For community detection on the whole networks, 20 processors were made available to R, but the methods were not parallelizable. For community detection using `SONNET`, all the $(r+1)s$ community detections were parallelized over 20 processors using the 'mclapply' function from the package 'parallel'. For `GALE`, all the p community detections were parallelized over 20 processors. Algorithm `GreedyMatch` was used for all the label matchings inside `SONNET` and to compute the error rates.

DBLP Four-Area Network Digital Bibliography & Library Project (DBLP) (`https://dblp.org/`) is a computer science bibliography website, jointly maintained by Schloss Dagstuhl - Leibniz Center for Informatics and the University of Trier. The website hosts over two million articles. Gao et al. (2009) and Ji et al. (2010) extracted a connected subset of the DBLP data, containing bibliographical records from four research areas related to data mining: database, data mining, information retrieval, and artificial intelligence. The original four-area dataset consists of 14376 papers written by 14475 authors, and presented at 20 conferences. However, the ground truth is available for 4057 authors with 14328 papers, presented in all the 20 con-

ferences. We use a version of the data that has the true community labels of the nodes so that the error rate can be computed.

Twitch Gamers Social Network The authors collected information of 168114 Twitch users in Spring 2018 using public application programming interface (API). The users are connected by an edge if they have mutual followers. There are 6 node features available for each user — explicit content identification, user language, user lifetime, dead account status, affiliate status, and view count. We performed community detection on a subnetwork of size $n = 32407$ to predict the user languages. To form the subnetwork, we obtained the largest connected component after removing all the users who broadcast in English, or have a dead account, or have view count or lifetime below their corresponding 10th percentiles. The subnetwork consists of $K = 20$ language communities. The ground truth communities are very unbalanced with their frequency distribution reported in Figure 3. We applied spectral clustering with row-normalization of the leading eigen vectors (`SC+RN`) on the entire network and `SONNET` with `SC+RN` and computed the error rates using the ground truth communities. We selected the parameters s and o for `SONNET` using the parameter selection method described in Section 3.3 for different values of the computation constraint q (as in (3.10)).

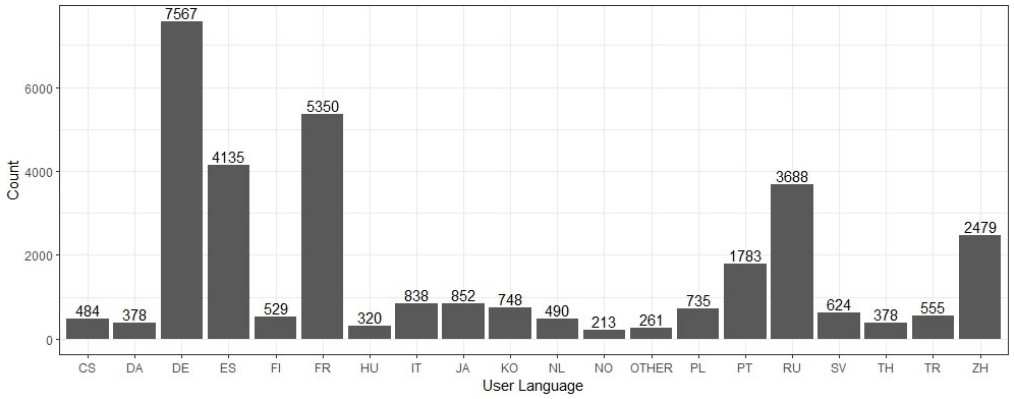

Figure 3: Frequency distribution of the ground truth communities of Twitch gamers subnetwork

Bibliography

Gao, J., Liang, F., Fan, W., Sun, Y., and Han, J. (2009), "Graph-based Consensus Maximization among Multiple Supervised and Unsupervised Models," in *Advances in Neural Information Processing Systems 22*, pp. 585–593.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010), "Graph Regularized Transductive Classification on Heterogeneous Information Networks," in *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg: Springer, pp. 570–586.

Lei, J. and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *Annals of Statistics*, 43, 215–237.

Mukherjee, S. S., Sarkar, P., and Bickel, P. J. (2021), "Two Provably Consistent Divide-and-Conquer Clustering Algorithms for Large Networks," *Proceedings of the National Academy of Sciences*, 118, e2100482118.

Serfling, R. J. (1974), "Probability Inequalities for the Sum in Sampling without Replacement," *The Annals of Statistics*, 2, 39 – 48.