Biostatistics (2023), **0**, 0, *pp.* 1–**??**
doi:10.1093/biostatistics/output

Supplementary Materials for "Identifying covariate-related subnetworks for whole-brain connectome analysis"

Shuo Chen[1,2,*], Yuan Zhang[3], Qiong Wu[4], Chuan Bi[2], Peter Kochunov[2], L.Elliot Hong[2],

[1]*Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD 21201, USA*

[2]*Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD 21201, USA*

[3]*Department of Statistics, Ohio State University, Columbus, OH 43210, USA*

[4]*Department of Biostatistics, Epidemiology, and Informatics, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

shuochen@som.umaryland.edu

1. Proof for Lemma 2.1

Proof of Lemma 2.1. We discuss the two cases.

- We first present the proof for the $H_{G;0}$ case. By Bernstein's inequality, for any $v \in \{v_0, \ldots, n\}$ and any $V \subseteq [n]$ satisfying $|V| = v$, we have

$$\mathbb{P}\left(\left|v^{-2}\sum_{i,j\in V}(A_{ij}-p)\right| > \gamma - p\right) \leqslant 2\exp\left(-\frac{v^4(\gamma-p)^2}{2[v^2+\frac{1}{3}v^2(\gamma-p)]}\right)$$
$$= 2\exp\left(-\left\{\frac{2}{(\gamma-p)^2}+\frac{2}{3(\gamma-p)}\right\}^{-1}\cdot v^2\right)$$

Therefore by a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\cup_{V\subseteq[n]:v=|V|\geqslant v_0}\left\{\left|v^{-2}\sum_{i,j\in V}(A_{ij}-p)\right|>\gamma-p\right\}\right)\\
\leqslant&\sum_{v=v_0}^{n}\binom{n}{v}\cdot 2\exp\left(-\left\{\frac{2}{(\gamma-p)^2}+\frac{2}{3(\gamma-p)}\right\}^{-1}\cdot v^2\right)\\
\leqslant&\sum_{v=v_0}^{n}2\exp\left(-\left\{\frac{2}{(\gamma-p)^2}+\frac{2}{3(\gamma-p)}\right\}^{-1}\cdot v^2+v\log n\right)\\
\leqslant&\sum_{v=v_0}^{n}2\exp\left(-\left\{\frac{4}{(\gamma-p)^2}+\frac{4}{3(\gamma-p)}\right\}^{-1}\cdot v^2\right)\\
\leqslant&2n\cdot\exp\left(-\left\{\frac{4}{(\gamma-p)^2}+\frac{4}{3(\gamma-p)}\right\}^{-1}\cdot v^2\right)
\end{aligned}$$

- Now we prove for the $H_{G;a}$ case. As both q and γ are in the range of $(0,1)$ $q-\gamma<3$ which ensures the positivity. The strategy is to simply consider G_c and show that with high probability, G_c would form a γ-quasi clique in $G[r]$. We have

$$\begin{aligned}
&\mathbb{P}\left\{\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}(G[r])_{ij}\geqslant\gamma|H_{G;a}\right\}\\
=&\mathbb{P}\left\{\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}\{(G[r])_{ij}-q\}\geqslant\gamma-q|H_{G;a}\right\}\\
\geqslant&1-\mathbb{P}\left\{\left|\binom{|G_c|}{2}^{-1}\sum_{i,j:(i,j)\in E(G_c)}\{(G[r])_{ij}-q\}\right|\geqslant q-\gamma|H_{G;a}\right\}\\
\geqslant&1-\exp\left\{-\frac{\frac{1}{2}(q-\gamma)^2\binom{|G_c|}{2}^2}{\binom{|G_c|}{2}+\frac{1}{3}(q-\gamma)\binom{|G_c|}{2}}\right\}\\
=&1-\exp\left\{-\frac{\frac{1}{2}(q-\gamma)^2\binom{|G_c|}{2}}{1+(q-\gamma)/3}\right\}
\end{aligned}$$

□

2. THEORETICAL RESULTS

In this subsection, we present theoretical guarantees for SICERS regarding covariate-related subnetwork detection and inference. Because our multiple testing procedure depends critically

on the correctness of the subnetwork detection by criterion (2.1), we first show a theoretical guarantee of the correctness of our optimization of (2.1). Our objective function with a ℓ_0 graph norm shrinkage criterion is:

$$\arg\max_{G,\tilde{C}} \log ||\mathbf{U}||_1 - \lambda_0 \log ||\mathbf{U}||_0. \tag{2.1}$$

THEOREM 2.1 (Optimality of subnetwork detection by (2.1)) Let C^* be the true number of subnetworks and $\boldsymbol{\pi} = (\pi_0, \pi_1, ..., \pi_{C^*}, \pi_{C^*+1})$ a vector of probabilities with $\|\boldsymbol{\pi}\|_1 = 1$, such that the membership of nodes toward subnetworks is generated by a multinomial distribution with parameter $\boldsymbol{\pi}$. Suppose that the tuning parameter is set to be $\lambda_0 \in (0, 1)$, and assume

$$\frac{\mu_0}{\mu_1} < \begin{cases} \frac{(C^*)^{\lambda_0}-1}{C^*-1} & \text{if } C^* \geqslant 2, \\ \lambda_0 & \text{if } C^* = 1. \end{cases} \tag{2.2}$$

Then, asymptotically, criterion (2.1) is uniquely optimized by $C = C^*$ and $G_c = G_c^*$ for all $c = 1, \ldots, C^*$.

Theorem 2.1 ensures that by optimizing (2.1), we can learn the correct number of subnetworks. This optimization is combinatorial and difficult to carry out in practice, but Theorem 2.1 suggests that criterion (2.1) can also be used for model selection when combined with efficient subnetwork estimation procedures for each candidate C. In view of this, next we present a theoretical guarantee of a computationally efficient estimation procedure for subnetwork detection under $C = C^*$. Let us define some notation. Recall the definitions of μ_0, μ_1, and define $\sigma_0^2 = \text{var}(w_{ij}|\delta_{ij} = 0)$ and $\sigma_1^2 = \text{var}(w_{ij}|\delta_{ij} = 1)$. Let $\boldsymbol{P} = \mathbb{E}[\boldsymbol{W}|G] = \boldsymbol{\Theta\Omega\Theta}^T$ denote the expectation matrix, where $\boldsymbol{\Theta} \in \{0, 1\}^{n\times(C+1)}$ is a membership matrix in which each row contains exactly one value of 1 with all other values set to 0. Here, $\boldsymbol{\Theta}_{i,(C+1)} = 1$ means that node i is a singleton node outside the subnetwork structure.

THEOREM 2.2 (Consistency of spectral estimation if $C = C^*$) Assume that $\text{rank}(\boldsymbol{P}) = C^* + 1$, and denote its smallest absolute nonzero eigenvalue by ξ_n. Assume $(\mu_1 \vee \sigma_1^2 \vee \sigma_0^2) \leqslant \alpha_n$ for

$\alpha_n \geqslant c_0 \log n/n$ and $c_0 > 0$. Then, if $(2+\varepsilon)\frac{(C+1)n\alpha_n}{\xi_n^2} < \tau$ for some $\tau, \varepsilon > 0$, the output $\hat{\mathbf{\Theta}}_{C^*}$ from the spectral estimation is consistent up to permutation. Equivalently, if $\hat{V}_c$ is the estimated node set for subgraph G_c, $c = 1, ..., C^*$, then $\hat{V}_c \cap V_c$ is the set in V_c for which the assignment of nodes can be guaranteed and, with probability at least $1 - n^{-1}$, up to permutation, we have

$$\sum_{c=1}^{C} \left[1 - \frac{\left|\hat{V}_c \cap V_c\right|}{|V_c|} \right] \leqslant \tau^{-1}(2+\varepsilon)\frac{Cn\alpha_n}{\xi_n^2}.$$

Theorems 2.1 and 2.2 provide two important results: the optimality of determining the number of subnetworks and the consistency of network recovery using the proposed algorithms. The assumptions of Theorems 2.1 and 2.2 involve the imbalanced distributions of subnetwork sizes, signal-to-noise ratio between within- and between-subnetwork edges, and overall sparsity of the graph.

THEOREM 2.3 Under the conditions of Theorem 2.1 and Lemma 2.1, Algorithm 2 is consistent, in the sense that

- when $H_{G;0}$ is true, we have $\mathbb{P}(\hat{C} = 0) \to 1$ and Algorithm 2 will not be executed;
- when $H_{G:a}$ is true, in Algorithm 2, with probability tending to 1, we would reject $H_{G;0}$ using each $\hat{G}_c$.

Proof of Theorem 2.1. We prove the population version, such that $\mathbf{W}$ satisfying $w_{ij} = \mu_1$ for $\delta_{ij} = 1$ and $w_{ij} = \mu_0$ for $\delta_{ij} = 0$. Denote $\mathbf{U}_{\mathbf{C}}^*$ as the matrix under true network structure G_C^*, i.e., $\mathbf{U}_C^* = \mathbf{W} * G_C^*$, and $\hat{\mathbf{U}}_C$ related with the optimized network structure under C, i.e., $\hat{\mathbf{U}}_C = \mathbf{W} * \hat{G}_C$.

For each $C \neq C^*$, let x be the number of corresponding edges with $\{\hat{\mathbf{U}}_C\}_{ij} = \mu_1$ and y to be $\{\hat{\mathbf{U}}_C\}_{ij} = \mu_1$. In other words, $x = \|\hat{\mathbf{U}}_C * G_C^*\|_0$ and $y = \|\hat{\mathbf{U}}_C\|_0 - \|\hat{\mathbf{U}}_C * G_C^*\|_0$. Then, the

objective function (10) takes value:

$$\begin{aligned} J_{\hat{\mathbf{U}}_C} &= \log \|\hat{\mathbf{U}}_C\|_1 - \lambda_0 \log \|\hat{\mathbf{U}}_C\|_0 \\ &= \log \frac{x\mu_1 + y\mu_0}{(x+y)^{\lambda_0}}. \end{aligned}$$

On the other hand, under true network structure G_C^*, the objective function (10):

$$\begin{aligned} J_{\mathbf{U}_C^*} &= \log \|\mathbf{U}_C^*\|_1 - \lambda_0 \log \|\mathbf{U}_C^*\|_0 \\ &= \log \frac{\|\mathbf{U}_C^*\|_0 \mu_1}{\|\mathbf{U}_C^*\|_0^{\lambda_0}} \geqslant \log \frac{x\mu_1}{x^{\lambda_0}} = J_{\hat{\mathbf{U}}_{C^* G_C^*}}, \end{aligned}$$

since the right-hand side is increasing in x for $lambda_0 \in (0,1)$, and $x \leqslant \mathbf{U}_C^*\|_0$ by definition.

Hence, to show our criterion (10) is optimized by $C = C^*$ and $G_c = G_c^*$ for all $c = 1, \ldots, C^*$, it suffices to have

$$\begin{aligned} J_{\hat{\mathbf{U}}_C} < J_{\hat{\mathbf{U}}_{C^* G_C^*}} &\iff \frac{x\mu_1 + y\mu_0}{(x+y)^{\lambda_0}} \leqslant \frac{x\mu_1}{x^{\lambda_0}} \\ &\iff \frac{\mu_0}{\mu_1} < \left[\left(1 + \frac{y}{x}\right)^{\lambda_0} - 1 \right] \frac{x}{y}, \end{aligned} \tag{2.3}$$

for each $C \neq C^*$ and $\hat{\mathbf{U}}_C$. Let $h(t) = \left[(1+t)^{\lambda_0} - 1\right] \frac{1}{t}$, then, $h'(t) = \frac{1}{t^2}\left[1 - \frac{(1-\lambda_0)t+1}{(1+t)^{1-\lambda_0}}\right]$. Since $(1+t)^a < 1 + at$ for $a \in (0,1)$ and $t > 0$, $h'(t)$ is negative and $h(t)$ is decreasing for all $t > 0$.

Therefore, it suffices to have

$$\frac{\mu_0}{\mu_1} < \left[\left(1 + \sup_{x,y} \frac{y}{x}\right)^{\lambda_0} - 1 \right] \frac{1}{\sup_{x,y} y/x}.$$

For each block $\hat{G}_c$, $c \in \{1, ..., C\}$, the nodes $\hat{V}_c$ of $\hat{G}_c$ are possible to have true memberships of at most C^* communities. Then, the number of edges in $\hat{G}_c$ with edge weight μ_1 would satisfy

$$\binom{g_1}{2} + \binom{g_2}{2} + ... + \binom{g_{C^*}}{2} \text{ with } g_1 + g_2 + ... + g_{C^*} = |\hat{V}_c|$$

where g_c is the number of nodes from the true community G_c^*. Consider a sufficiently large graph and the numbers of nodes change continuously, we have

$$\frac{\binom{g_1}{2} + \binom{g_2}{2} + ... + \binom{g_{C^*}}{2}}{\binom{|\hat{V}_c|}{2}} \geqslant C^*.$$

Therefore, for $C* \geqslant 2$, $y/x \leqslant C^* - 1$ and for $C^* = 1$, $y/x \leqslant 1$. Hence, the claim is true. □

Proof of Theorem 2.2. It suffices to show that consistent results are guaranteed for spectral clustering in our setting of a continuous stochastic block model. The proof of theorem 3.1 in [6] can be easily extended to a weighted case using continuous versions of Bernstein inequality and Chernoff bounds.

To bound light pairs, $u_{ij} = x_i y_j \mathbf{1}(|x_i y_j| \leqslant \sqrt{d}/n) + x_j y_i \mathbf{1}(|x_j y_i| \leqslant \sqrt{d}/n)$, then $|u_{ij}| \leqslant 2\sqrt{d}/n$, and $x^T W' y$ can be written as

$$\sum_{1 \leqslant i < j \leqslant n} w'_{ij} u_{ij}.$$

Then, for zero-mean independent random variables, apply Bernstein inequality,

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{i<j} w'_{ij} u_{ij}\right| \geqslant c_0\sqrt{d}\right] &\leqslant 2\exp\left(-\frac{\frac{1}{2}c_0^2 d}{\sum_{i<j}\sigma_{ij}^2 u_{ij}^2 + \frac{1}{3}\frac{2\sqrt{d}}{n}c_0\sqrt{d})}\right) \\
&\leqslant 2\exp\left(-\frac{\frac{1}{2}c_0^2 d}{\sigma_{\max}^2 \sum_{i<j} u_{ij}^2 + \frac{2c_0}{3}\frac{d}{n}}\right) \\
&\leqslant 2\exp\left(-\frac{c_0^2}{4 + \frac{4c_0}{3}}n\right).
\end{aligned}$$

In bounding heavy pairs, let $e(I, J)$ be the summation of edge weights in node sets I and J: $e(I, J) = \sum_{(i,j)\in s(I,J)} w_{ij}$. Define $\mu(I, J) = \mathbb{E}e(I, J), \overline{\mu}(I, J) = p_{\max}|I||J|$. We could obtain continuous versions of Lemma 4.1 and 4.2 in the supplementary material of [6].

Using Bernstein inequality:

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^n w_{ij} \geqslant c_1 d\right) &\leqslant \mathbb{P}\left(\sum_{j=1}^n w'_{ij} \geqslant (c_1 - 1)d\right) \leqslant \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d^2}{\sum_{j=1}^n \sigma_{ij}^2 + \frac{1}{3}(c_1-1)d}\right] \\
&\leqslant \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d^2}{n\sigma_{\max}^2 + \frac{1}{3}(c_1-1)d}\right] \leqslant \exp\left[-\frac{\frac{1}{2}(c_1-1)^2 d}{1 + \frac{1}{3}(c_1-1)}\right] \leqslant n^{-\frac{3c_0(c_1-1)^2}{2c_1+4}}
\end{aligned}$$

We have for $c_0 > 0$, there exists constant $c_1 = c_1(c_0)$ such that with probability at least $1 - n^{-c_0}$, $\sum_{j=1}^n w_{ij} \leqslant c_1 d$.

From Chernoff Bound:

$$\begin{aligned}\mathbb{P}[e(I,J) \geqslant k\bar{\mu}(I,J)] &= \mathbb{P}\left[\sum_{(i,j)\in s(I,J)} w_{ij} \geqslant k\bar{\mu}(I,J)\right] \\ &\leqslant \exp(-\bar{\mu}(I,J)(k\ln k - (k-1))) \\ &\leqslant \exp\left[-\frac{1}{2}(k\ln k)\bar{\mu}\right]\end{aligned}$$

the lemma 4.2 is true from exactly the same calculations.

Hence, our claim is true with stated assumptions from Theorem 3.1 of [6]. □ □

Proof of Theorem 2.3. We discuss the two cases.

- When $H_{G;0}$ is true, according Theorem 2.1, we have $\mathbb{P}(\hat{C} = 0) \to 1$.

- When $H_{G:a}$ is true, assuming $\min_{c=1,\ldots,C^*} |G_c| \geqslant c_0\sqrt{n}$ and the conditions of Theorem 2.1 hold, by Theorem 2.1, we have

  $$\hat{C} \xrightarrow{p} C^*; \quad \text{and} \quad \{\hat{G}_c\}_{c=1,\ldots,\hat{C}} \xrightarrow{p} \{G_c\}_{c=1,\ldots,C^*}$$

  Therefore,

  $$\hat{\mu}_0 \xrightarrow{p} \mu_0; \quad \text{and} \quad \hat{\mu}_1 \xrightarrow{p} \mu_1; \quad \text{and} \quad \hat{r} \xrightarrow{p} (\mu_0+\mu_1)/2$$

  which further yields

  $$\hat{p} \xrightarrow{p} p(r); \quad \text{and} \quad \hat{q} \xrightarrow{p} q(r); \quad \text{and} \quad \hat{\gamma} \xrightarrow{p} \gamma(r)$$

  Consequently, with probability tending to 1, we have $\hat{\gamma} - \hat{p}$ bounded away from zero by a constant gap, therefore the empirical p-value converges in probability to zero.

□

Proof of Theorem 2.4. We first prove (2.9). By the definition of most liberal multiple testing, we have

$$M_0 = \binom{n}{2} \cdot \alpha \tag{2.4}$$

Now since our multiple testing algorithms do not reject any individual null hypotheses $H_{(i,j)}$ where $(i,j) \notin \cup_{c=1}^{C} V_c \times V_c$, we immediately have

$$M_1 = \mathbb{E}\left[\sum_{c=1}^{C} \binom{\hat{n}_c}{2} \cdot \alpha\right] \tag{2.5}$$

Since

$$\begin{aligned}
\frac{\sum_{c=1}^{C} \binom{\hat{n}_c}{2} \cdot \alpha}{\binom{n}{2} \cdot \alpha} &= \sum_{c=1}^{C} \frac{\hat{n}_c(\hat{n}_c - 1)}{n(n-1)} \\
&= \sum_{c=1}^{C} \left(\frac{n_c}{n} + \frac{\hat{n}_c - n_c}{n}\right)\left(\frac{n_c}{n} + \frac{\hat{n}_c - n_c}{n} - \frac{1}{n}\right) / \left(1 - \frac{1}{n}\right) \\
&\leqslant \frac{n_c^2}{n^2} + \frac{2n_c|\hat{n}_c - n_c|}{n^2} + \frac{(\hat{n}_c - n_c)^2}{n^2} \\
&= \frac{n_c^2}{n^2}\left\{1 + \frac{2|\hat{n}_c - n_c|}{n_c} + \frac{(\hat{n}_c - n_c)^2}{n_c^2}\right\} \\
&\leqslant p_0^2\left(1 + 4\sum_{c=1}^{C} \frac{|\hat{n}_c - n_c|}{n_c}\right)
\end{aligned} \tag{2.6}$$

Then combining this with Lemma 2.2 proves (2.9).

Now we prove (2.10). Define the power of an individual test to be β. By definition, we have

$$N_0 = \sum_{c=1}^{C} \binom{n_c}{2} \gamma_c \beta$$

Now we consider N_1. For each $c = 1, \ldots, C$, the estimated and true community c share $n_c - n_{\text{misclass};c}$ nodes. If all the individual test outside $V_c \cap \hat{V}_c$ accept their null hypotheses, then the contribution of power from the estimated class c would be

$$\beta \cdot \left\{\binom{n_c}{2}\gamma_c - (n_c - n_{\text{misclass};c}/2) \cdot n_{\text{misclass};c}\right\}$$

In order to lower bound N_1/N_0, observe that

$$\begin{aligned}
\frac{\beta \cdot \left\{\binom{n_c}{2}\gamma_c - (n_c - n_{\text{misclass};c}/2) \cdot n_{\text{misclass};c}\right\}}{\binom{n_c}{2}\gamma_c\beta} &= 1 - \frac{2(n_c - n_{\text{misclass};c}/2) \cdot n_{\text{misclass};c}}{n_c(n_c - 1)} \\
&= 1 - \frac{(2 - n_{\text{misclass};c}/n_c) \cdot n_{\text{misclass};c}/n_c}{(1 - n_c^{-1})} \\
&\geqslant 1 - \frac{4n_{\text{misclass};c}}{n_c}
\end{aligned} \tag{2.7}$$

Therefore

$$\frac{\sum_{c=1}^{C} \beta \cdot \left\{ \binom{n_c}{2} \gamma_c - (n_c - n_{\text{misclass};c}/2) \cdot n_{\text{misclass};c} \right\}}{\sum_{c=1}^{C} \binom{n_c}{2} \gamma_c \beta} \geqslant 1 - \sum_{c=1}^{C} \frac{4 n_{\text{misclass};c}}{n_c} \tag{2.8}$$

where we used the following basic fact that

$$\frac{a_c}{b_c} \geqslant 1 - e_c, \forall c = 1, \ldots, C \quad \Rightarrow \quad \frac{\sum_{c=1}^{C} a_c}{\sum_{c=1}^{C} b_c} \geqslant 1 - \sum_{c=1}^{C} e_c$$

where a, b, e are all positive numbers and $e \in (0, 1)$. Finally taking an expectation on both sides of (2.8) completes the proof of (2.10). □

Next, we compare the accuracy of covariate-related subnetwork-wise and edge-wise inference. The following theorem compares the false positive error rates and sensitivity of SICERS vs. edge-wise inference with a universal cut-off. We assume that all edges in a significant subnetwork are covariate-correlated.

THEOREM 2.4 (Sensitivity and false positive error rate) Define

$$p_0 = \sup_{c=1,\ldots,C} |V_c|/n,$$

and denote the expected numbers of false positive edges across multiple tests $H_{(1,2)}, \ldots, H_{n-1,n}$ based on subnetwork-wise and edge-wise inference, using α as a universal threshold, by M_1 and M_0, respectively. Under the conditions of Theorem 2.1, we have

$$\frac{M_1}{M_0} \leqslant p_0^2 \left\{ C + 4\tau^{-1}(2+\epsilon) \frac{Cnd}{\xi_n^2} \right\}. + n^{-1} \tag{2.9}$$

On the other hand, suppose that within each subnetwork G_c, a proportion of γ_c individual alternative hypotheses are true. Denote the expected numbers of true positive edges based on subnetwork-wise and edge-wise inference (using α as a universal threshold) by N_1 and N_0, respectively. Then, we have

$$\frac{N_1}{N_0} \geqslant 1 - 4\tau^{-1}(2+\epsilon) \frac{Cnd}{\xi_n^2}. - n^{-1} \tag{2.10}$$

Theorem 2.4 theoretically justifies the significant improvement in subnetwork-wise inference (SICERS) compared with edge-wise inference with respect to false positive error rate and sensitivity (power) in the context of multiple testing. The theoretical advantage of our method is confirmed by simulations and data examples, as demonstrated in the next section.

3. Implementing the objective function (2.1) for subnetwork extraction

We implement the objective function (2.1) for subnetwork extraction by Algorithm 1 in the main text. Specifically, we optimize (2.1) with given C, and then select the optimal C. Here, we focus on optimizing (2.1) with a given C. The objective

$$\begin{aligned} &\underset{\hat{\mathbf{U}}=\cup_{c=1}^{C}\hat{\mathbf{U}}_c}{\arg\max}\ \log||\hat{\mathbf{U}}||_1-\lambda_0\log||\hat{\mathbf{U}}||_0 \\ =&\underset{\hat{\mathbf{U}}=\cup_{c=1}^{C}\hat{\mathbf{U}}_c}{\arg\max}\ \log\Big(\frac{||\hat{\mathbf{U}}||_1}{||\hat{\mathbf{U}}||_0^{\lambda_0}}\Big)\doteq\underset{\hat{\mathbf{U}}=\cup_{c=1}^{C}\hat{\mathbf{U}}_c}{\arg\max}\ f(\hat{\mathbf{U}}) \end{aligned} \tag{3.11}$$

We start by setting $\lambda_0=0.5$ reflecting balanced covering quality and quantity of true positive edges, and the objective function (3.11) then becomes the well-known problem of k dense subgraph discovery, where $f(\cdot)$ is the density function. The problem has been solved in polynomial time by Goldberg's min-cut algorithm (5) and a greedy algorithm with 1/2 approximation by [2]. In addition, the default topological community structure can be considered as quasi-cliques and the problem can be solved by additive approximation algorithms and local-search heuristics (8). Alternatively, with the mild spatially-invariant assumptions that $\frac{E(w_{ij}|e_{ij})\in G_c}{|E_c|}=\rho_1,\forall c,\ 0\leqslant c\leqslant C$, and $\frac{E(w_{ij}|i\in V_c,j\in V_{c'})}{|V_c||V_{c'}|}=\rho_0,\forall c,\ 0\leqslant c\leqslant C$ the primary objective function is equivalent to

$$\begin{aligned}
&\underset{\hat{\mathbf{U}}=\cup_{c=1}^{C}\hat{\mathbf{U}}_c}{\arg\min}\ \log\frac{\sum_{c=1}^{C}\sum_{i<j}(w_{ij}|e_{ij}\notin G_c)}{[\sum_{c=1}^{C}\sum_{i<j}I(e_{ij}\notin G_c)]}\\
\doteq &\underset{\hat{\mathbf{U}}=\cup_{c=1}^{C}\hat{\mathbf{U}}_c}{\arg\min}\ \log\sum_{c=1}^{C}\frac{\sum_{i<j}(w_{ij}|e_{ij}\notin G_c)}{|V_c|}, \text{with spatially invariant } \rho_0
\end{aligned} \tag{3.12}$$

Although the objective function (3.12) is not convex, the issue of local optima in the discrete optimization can be solved by restarting the algorithm several times with different initializations and/or through orthonormal transforms (7 and 1). The proposed algorithm may better extract multiple weighted dense subgraphs (with an unknown number and unknown sizes of dense subgraphs) than the existing algorithms of dense subgraph discovery (4). We then choose the optimal C^* by grid searching that maximizes the following criteria:

$$\underset{C^*}{\arg\max}\left(\frac{\sum_{c=1}^{C^*}\sum_{i<j}(w_{i,j}|e_{i,j}\in G_c)}{\sum_{c=1}^{C^*}|E_c|}\right)^{\lambda_0}\left(\sum_{c=1}^{C^*}\sum_{i<j}(w_{i,j}|e_{i,j}\in G_c)\right)^{1-\lambda_0}. \tag{3.13}$$

The criteria (3.13) can be directly derived from our primary objective function that

$$\log(\sum_{c=1}^{C^*}\sum_{i<j}(w_{i,j}|e_{i,j}\in G_c))-\lambda_0\log||\mathbf{U}||_0.$$

The first term in (3.13) indicates the 'quality' (the area density) of the extracted subgraphs, while the second term represents the 'quantity' of edges covered by the subgraphs. C^* is selected with optimal quality and quantity in terms of covering informative edges. λ_0 can be tuned to either extract subgraphs with higher area density (i.e. low false positive rates) or cover more high-weight edges using subgraphs with larger sizes (i.e. low false negative rates). In general, C^* selection is robust for λ_0 in the range of 0.4 to 0.7.

We can objectively select the optimal tuning parameter λ_0 based on the likelihood function of G^β. We obtain G^β by binarize $e_{ij}^\beta = I(w_{ij} > r)$. Then, we calculate the likelihood for G^β under $\hat{G}(\lambda_0) = \cup_{c=1}^{C}\hat{G}_c(\lambda_0)\cup\hat{G}_0$ vs. the null $G = G_0$ and integrate the cutoff r by a prior distribution

$g(r)$ (i.e., belief where r can be a good cut-off).

$$lr(\cup_{c=1}^{\hat{C}}\hat{G}_c(\lambda_0)||G) = \left\{\int_r \left[\sum_{i,j\in\hat{G}_c(\lambda_0)}\left(e_{ij}^{\beta}\log\frac{\pi_1}{\pi}+(1-e_{ij}^{\beta})\log\frac{(1-\pi_1)}{(1-\pi)}\right) + \sum_{i,j\notin\hat{G}_c(\lambda_0)}\left(e_{ij}^{\beta}\log\frac{\pi_0}{\pi}+(1-e_{ij}^{\beta})\log\frac{(1-\pi_0)}{(1-\pi)}\right)\right] g(r)dr\right\}. \quad (3.14)$$

where

$$\pi := \frac{\sum_{1\leqslant i<j\leqslant n} I(w_{ij}>\hat{r})}{\binom{n}{2}}, \pi_1 := \frac{\sum_{(i,j)\in\hat{G}_c} I(w_{ij}>\hat{r})}{\|\hat{G}_c\|_0}, \pi_0 := \frac{\sum_{(i,j)\notin\hat{G}_c} I(w_{ij}>\hat{r})}{\binom{n}{2}-\sum\|\hat{G}_c\|_0}$$

We select λ_0 that most deviates from the null that G^{β} is a random graph.

In summary, the above procedure can extract latent organized topological structures containing the most high-weight edges while controlling the sizes of the topological structures by ℓ_0 norm regularization. Furthermore, we have recently developed more flexible algorithms to extract subgraphs beyond the default community structure, for example, k-partite/rich club and interconnected induced subgraphs can be further detected based on detected quasi-cliques (3 and 9). These more sophisticated topological structures can further improve the objective function by preserving the high-weight edges inside of more parsimoniously-sized subgraphs.

4. FMRI DATA ACQUISITION AND PRE-PROCESSING

All participants provided written informed consent that had been approved by the University of Maryland Internal Review Board. All participants were evaluated using the Structured Clinical Interview for the DSM-IV diagnoses. We recruited medicated patients with an Axis I diagnosis of schizophrenia through the Maryland Psychiatric Research Center and neighboring mental-health clinics. We recruited control subjects, who did not have an Axis I psychiatric diagnosis, through media advertisements. Exclusion criteria included hypertension, hyperlipidemia, type 2 diabetes, heart disorders, and major neurological events, such as stroke or transient ischemic attack. Illicit substance and alcohol abuse and dependence were exclusion criteria. Data were acquired using a 3-T Siemens Trio scanner equipped with a 32-channel head coil at the University of Maryland Center for Brain Imaging Research. A T1-weighted structural image (MP-RAGE: 1 mm isotropic voxels, 256 x 256 mm FOV, TR/TE/TI = 1900/3.45/900ms) was acquired for anatomical reference. Fifteen minutes of rfMRI was collected on each subject. During the resting scans, subjects were given a simple instruction to rest and keep their eyes closed. Head motion was minimized using foam padding, foam molding, and tapes. RfMRI were acquired over 39 axial, interleaving slices using a gradient-echo EPI sequence (450 volumes, TE/TR = 27/2000 ms; flip angle = 90^o; FOV = 220x220 mm; image matrix = 128x128; in-plane resolution 1.72x1.72mm. Following the previously published procedures, data were preprocessed in AFNI and MATLAB (MathWorks, Inc., Natick, MA). Volumes were slice-timing aligned and motion corrected to the base volume that minimally deviated from other volumes using an AFNI built-in algorithm. After linear detrending of the time course of each voxel, volumes were spatially normalized and resampled to Talairach space at 3 mm^3, spatially smoothed (FWHM 6 mm), and temporally low-pass filtered (0.1 Hz). For functional connectivity analyses, the six rigid head-motion parameter time courses and the average time course in white matter were treated as nuisance covariates. A white matter mask was generated by segmenting the high-resolution anatomical images and

down-gridding the obtained white matter masks to the same resolution as the functional data. These nuisance covariates regress out fluctuations unlikely to be relevant to neuronal activity.

5. JUSTIFICATION OF USING p-VALUES

We use p-values as a measure to assess the association between the covariate and the connectome, thereby capturing the covariate-related subnetwork patterns in section 2.4. p-values are widely used in the analysis of genetic, genomic, neuroimaging, and other high-throughput data sets. Popular techniques, including false discovery rate (FDR), Manhattan plots in GWAS, NBS, and volcano plots, rely on p-values to identify high-throughput features that are associated with the covariate. Fig. 1 demonstrates that p-values can better discern $\beta_{ij} \neq 0$ vs $\beta_{ij} = 0$ than test statistics.

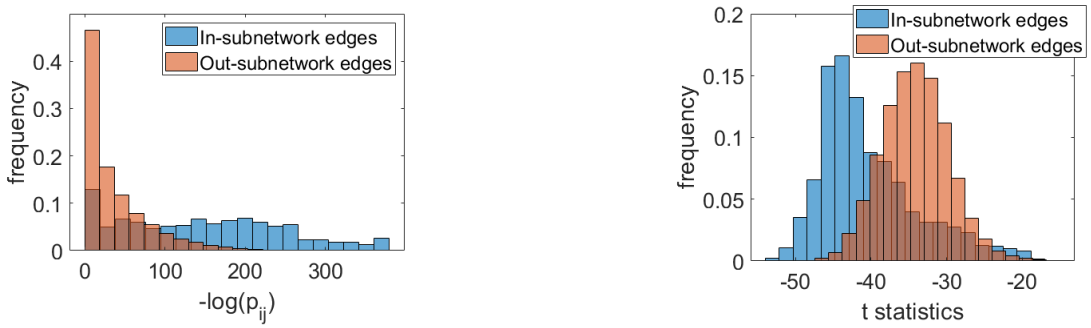

Fig. 1: Left panel demonstrates that $-\log$ of p-values can be used to capture the patterns of connectome associated with the covariate. $-\log$ of p-values can better capture the latent patterns with a larger range than test statistics in the left panel. The figure is based on connectome associated with the biological sex variable using Uk biobank data. See Fig. 4. See Fig. 4.

6. INFLUENCE OF THE COVARIATE-RELATED SUBNETWORK SIZE ON INFERENCE

In general, all network detection algorithms become less effective as the sizes of subnetworks decrease. We perform an additional simulation study to test the influence of subnetwork size based on effect size Cohen's $d = 0.5$) and sample size $S = 240$ for 100 repeated simulations.

The results suggest that all subnetwork extraction methods require the size of a subnetwork with $V_c \geqslant 20$ to achieve valid power and FPR.

	size	30	20	10	5
SICERS	Power	1(0)	1(0)	0.125(0.35)	0(0)
	FPR	0.2(0.17)	0.4(0.19)	0.95(0.04)	1(0)
Louvain	Power	1(0)	1(0)	0.5(0.53)	0(0)
	FPR	0.4(0.14)	0.8(0.06)	0.98(0.05)	1(0)
Dense	Power	1(0)	1(0)	0(0)	0(0)
	FPR	0.5(0.12)	0.6(0.21)	1(0)	1(0)
NBS	Power	0(0)	0(0)	0(0)	0(0)
	FPR	1(0)	1(0)	1(0)	1(0)

Table 1: Network-level power and FPR for various network sizes. 30, 20, 10, and 5 are used as subnetwork sizes. The effect size (Cohen's d) is 0.5, sample size $S = 240$, and $\alpha = 0.05$. Values presented in the tables are mean values over 100 repeated simulations, where standard deviations are recorded in the parenthesis.

7. Impact of tuning parameter λ_0 on the objective function

To demonstrate varying λ_0 would affect the performance of the our objective function, we show the performance of subnetwork extraction with different values of λ_0, in Figure 2.

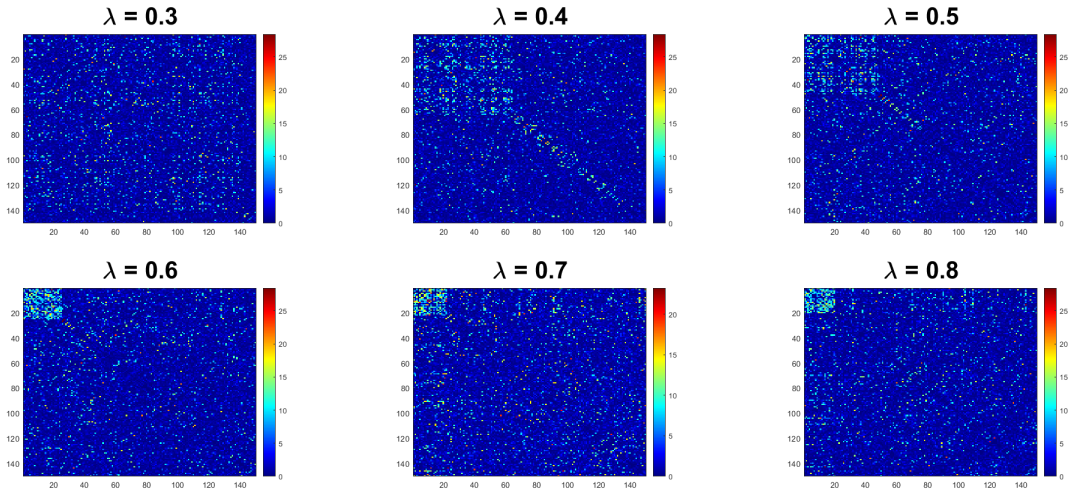

Fig. 2: Extracted subnetwork patterns using different values of λ_0, where the number of nodes for the ground truth subnetwork is 25. The optimal λ_0 is obtained around $\lambda_0 = 0.6$.

8. SUBNETWORK DETECTION FOR SPARSE NETWORKS

When the key assumption is violated and covariate-related subnetworks are sparse graphs, the SICERS and other network extraction algorithms may not identify any covariate-related subnetworks. Therefore, no false positive covariate-related subnetworks will be reported. Alternatively, we can use edge-level inference tools, including FDR and FWER, to identify individual edges. We further perform additional simulation analysis to examine whether SICERS can capture the sparse graphs. Table 2 summarizes the network-level results for sparse networks. We generated three sparse networks with $1\%, 2\%$, and 3% of edges that are associated with covariate. Because the covariate-related edges are randomly distributed and not included in any subnetworks, we consider the significant network findings as false positive findings (i.e., FPR). The results in Table 2 show that none of the used methods reports false positive subnetworks. This further suggests that the edge-level inference method (e.g., FWER) should be used for the scenario of sparse covariate-related graphs.

Percent. of Sig. Edges	1%	2%	3%
SICERS (Network-FPR)	0(0)	0(0)	0(0)
Louvain (Network-FPR)	0(0)	0(0)	0(0)
greedy (Network-FPR)	0(0)	0(0)	0(0)
NBS (Network-FPR)	0(0)	0(0)	0(0)

Table 2: Network-level results for sparse covariate-related graphs. We set $1\%, 2\%$, and 3% of 4950 edges in G that are associated with the covariate. Then, we apply the network-inference methods to identify covariate-related subnetworks. We summarize the means (standard deviations) of network-level 'FPR'. None of these methods reports significant subnetworks.

9. COVARIATE-RELATED SUBNETWORKS: ADDITIONAL DEMONSTRATION

We apply SICERS to another data example of 20,100 participants from UK biobank data. The two covariates are age and sex. First, we investigate the brain connectome decline patterns associated with age. The results are illustrated in Fig. 3. We see that there are two brain subnetworks that are associated with age-related decline, where the subnetworks mainly consist of the following

brain regions: cluster 1 (Insular Gyrus, Cingulate Gyrus, Inferior Parietal Lobule, and Superior Temporal Gyrus), cluster 2(Basal Ganglia and Thalamus). Secondly, we test the difference between males vs. females, shown in Fig. 4. As a result, a dense covariate-related subnetwork is identified, where the major brain regions associated with the subnetwork are temporal, insular lobes. Additional subregions belonging to the frontal and parietal lobes are also included in the subnetwork. The subnetwork demonstrates systematical hypo-connections in FC among female subjects.

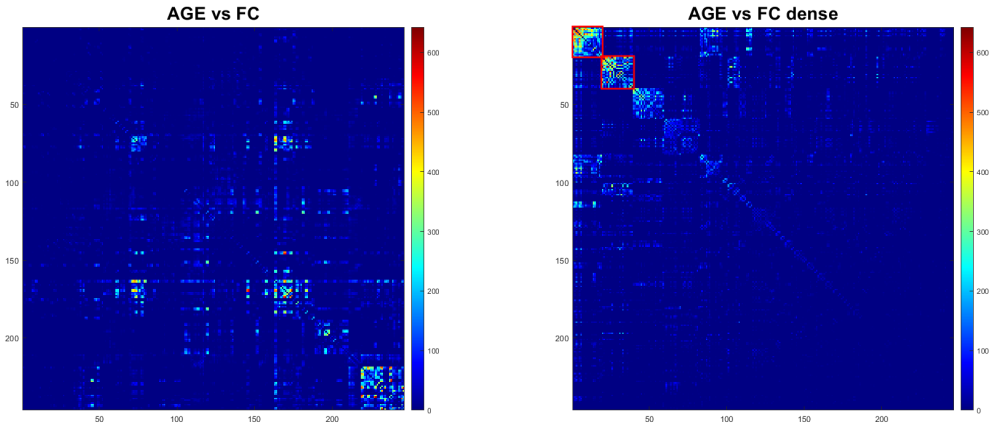

Fig. 3: Age-related subnetworks. Left: functional brain network representing the $-\log(p)$ values with regions in Atlas order; right: we reorder the regions based on identified dense subnetworks.

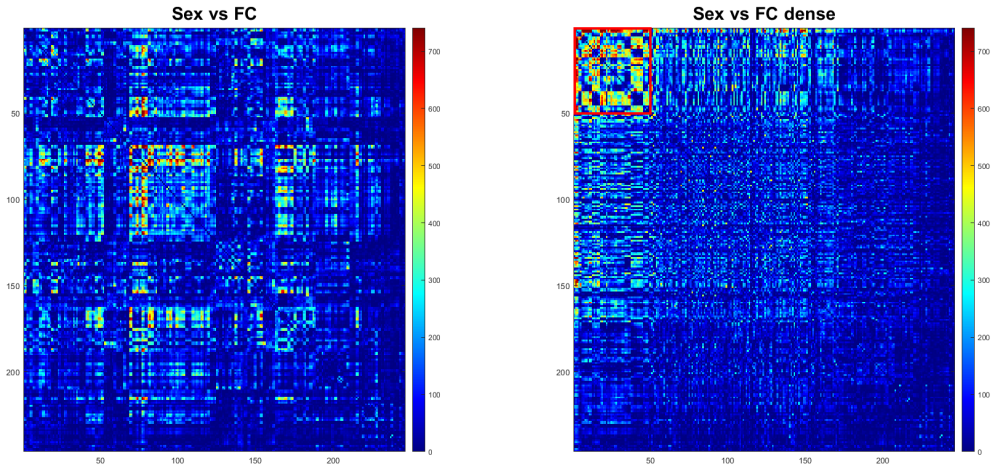

Fig. 4: Sex-related subnetworks. Left: functional brain network representing the $-\log(p)$ values with regions in Atlas order; right: we reorder the regions based on identified dense subnetworks.

10. 3D DEMONSTRATION OF SCHIZOPHRENIA-RELATED SUBNETWORKS

11. TABLES OF BRAIN REGIONS

In the following tables, we list the region names and coordinates of subnetworks from D^1 and D^2.

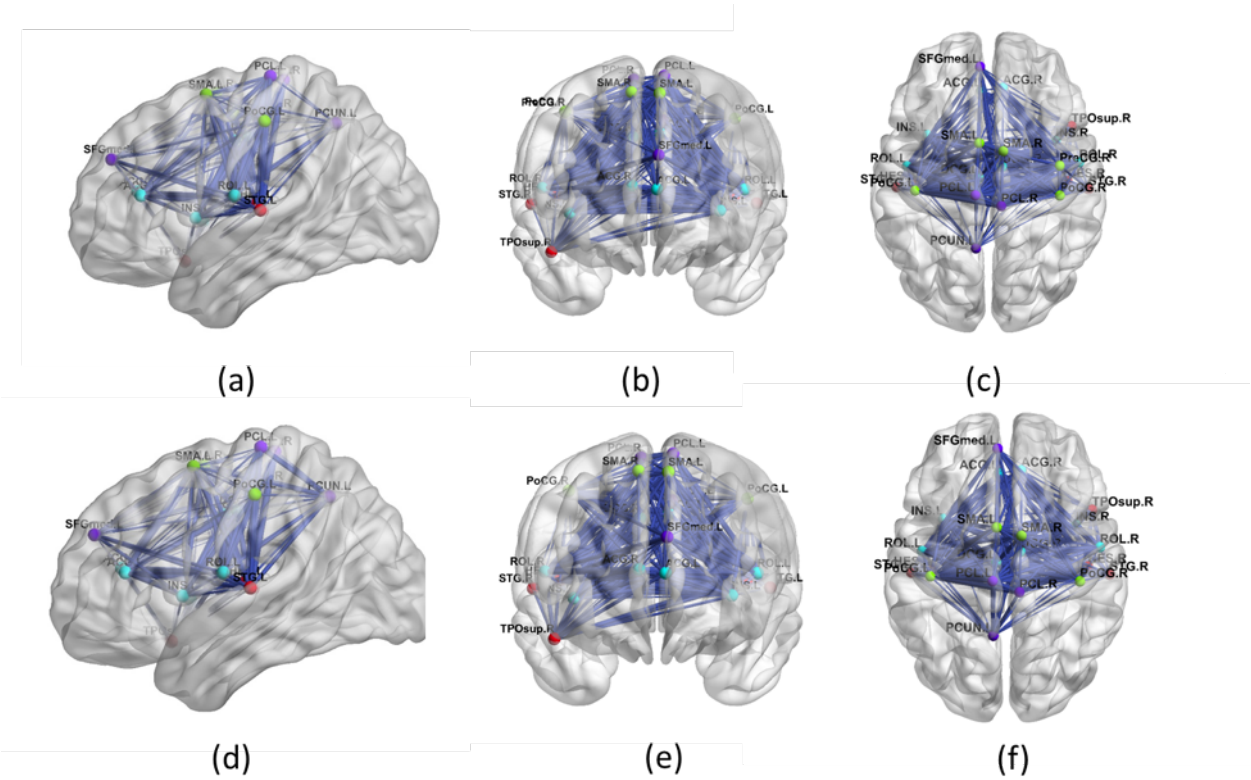

Fig. 5: The edges in the subnetwork using 3D demonstration for data set 1 (a)–(c) and data set 2 (d)–(f). The line width is proportional to the effect size. The disease-relevant network involves the SN, part of the DMN, and part of the executive network, and more importantly, the interconnections among these three networks are revealed. Panels (e) and (f) show the 3D brain subnetwork for data set 2, which shows a highly replicable brain subnetwork as seen in data set 1 with one fewer brain region (precentral R).

References

Marianna Bolla. *Spectral clustering and biclustering: Learning large graphs and contingency tables*. John Wiley & Sons, 2013.

Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 84–95. Springer, 2000.

Table 3: Region names and coordinates in the subnetwork of D^1

Region Name	x	y	z
Precentral R	-39	-6	51
Rolandic Oper L	-47	-8	14
Rolandic Oper R	53	-6	15
Supp Motor Area L	-5	5	61
Supp Motor Area R	9	0	62
Frontal Sup Medial L	-5	49	31
Insula L	-35	10	3
Insula R	39	6	2
Cingulum Ant L	-4	35	14
Cingulum Ant R	8	37	16
Cingulum Mid L	-5	-15	42
Cingulum Mid R	8	-9	40
Postcentral L	-42	-23	49
Postcentral R	41	-25	53
Precuneus L	-7	-56	48
Paracentral Lobule L	-8	-25	70
Paracentral Lobule R	7	-32	68
Heschl L	-42	-19	10
Heschl R	46	-17	10
Temporal Sup L	-53	-21	7
Temporal Sup R	58	-22	7
Temporal Pole Sup R	48	15	-17

Shuo Chen, F DuBois Bowman, and Yishi Xing. Detecting and testing altered brain connectivity networks with k-partite network topology. *Computational Statistics & Data Analysis*, 2019.

Shuo Chen, Jian Kang, Yishi Xing, Yunpeng Zhao, and Donald K Milton. Estimating large covariance matrix with network topology for high-dimensional biomedical data. *Computational Statistics & Data Analysis*, 127:82–95, 2018.

Andrew V Goldberg. *Finding a maximum density subgraph.* University of California Berkeley, CA, 1984.

Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003.

Table 4: Region names and coordinates in the subnetwork of D^2

Region Name	x	y	z
Rolandic Oper L	-47	-8	14
Rolandic Oper R	53	-6	15
Supp Motor Area L	-5	5	61
Supp Motor Area R	9	0	62
Frontal Sup Medial L	-5	49	31
Insula L	-35	10	3
Insula R	39	6	2
Cingulum Ant L	-4	35	14
Cingulum Ant R	8	37	16
Cingulum Mid L	-5	-15	42
Cingulum Mid R	8	-9	40
Postcentral L	-42	-23	49
Postcentral R	41	-25	53
Precuneus L	-7	-56	48
Paracentral Lobule L	-8	-25	70
Paracentral Lobule R	7	-32	68
Heschl L	-42	-19	10
Heschl R	46	-17	10
Temporal Sup L	-53	-21	7
Temporal Sup R	58	-22	7
Temporal Pole Sup R	48	15	-17

Charalampos Tsourakakis, Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Maria Tsiarli. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 104–112. ACM, 2013.

Qiong Wu, Tianzhou Ma, Qingzhi Liu, Donald Milton, Yuan Zhang, and Shuo Chen. Extracting interconnected communities in gene co-expression networks. 2021.