Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model

Kamalika Chaudhuri KAMALIKA@CS.UCSD.EDU
Fan Chung FAN@CS.UCSD.EDU
Alexander Tsiatas ATSIATAS@CS.UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

In this paper, we examine a spectral clustering algorithm for similarity graphs drawn from a simple random graph model, where nodes are allowed to have varying degrees, and we provide theoretical bounds on its performance. The random graph model we study is the Extended Planted Partition (EPP) model, a variant of the classical planted partition model.

The standard approach to spectral clustering of graphs is to compute the bottom k singular vectors or eigenvectors of a suitable graph Laplacian, project the nodes of the graph onto these vectors, and then use an iterative clustering algorithm on the projected nodes. However a challenge with applying this approach to graphs generated from the EPP model is that unnormalized Laplacians do not work, and normalized Laplacians do not concentrate well when the graph has a number of low degree nodes.

We resolve this issue by introducing the notion of a degree-corrected graph Laplacian. For graphs with many low degree nodes, degree correction has a regularizing effect on the Laplacian. Our spectral clustering algorithm projects the nodes in the graph onto the bottom k right singular vectors of the degree-corrected random-walk Laplacian, and clusters the nodes in this subspace. We show guarantees on the performance of this algorithm, demonstrating that it outputs the correct partition under a wide range of parameter values. Unlike some previous work, our algorithm does not require access to any generative parameters of the model.

Keywords: Spectral clustering, unsupervised learning, normalized Laplacian

1. Introduction

Spectral clustering of similarity graphs is a fundamental tool in exploratory data analysis, which has enjoyed much empirical success (Shi and Malik, 2000; Ng et al., 2002; von Luxburg, 2007) in machine-learning. In this paper, we examine a spectral clustering algorithm for similarity graphs drawn from a simple random graph model, where nodes are allowed to have varying degrees, and we provide theoretical bounds on its performance. Such clustering problems arise in the context of partitioning social network graphs to reveal hidden communities, or partitioning communication networks to reveal groups of nodes that frequently communicate.

The random graph model we study is the Extended Planted Partition (EPP) model, a variant of the classical planted partition model. A graph $G = (V, E)$ generated from this model has a hidden partition $V_1, \ldots, V_k$, as well as a number d_u associated with each node u. If two nodes u and v

© 2012 K. Chaudhuri, F. Chung & A. Tsiatas.

lie in the same cluster V_i, then the edge (u, v) is present in G with probability $d_u p d_v$; otherwise, it is present with probability $d_u q d_v$. Given a graph G generated from this model, the Extended Planted Partition problem is to recover the hidden partition $V_1, \ldots, V_k$ without prior knolwedge of the specific model parameters.

The standard approach to spectral clustering of graphs is to compute the bottom k singular vectors or eigenvectors of a suitable graph Laplacian, project the nodes of the graph onto these vectors, and then use an iterative clustering algorithm on the projected nodes. This approach performs well when nodes from different clusters are well-separated when projected on to the bottom k singular space or eigenspace of the appropriate graph Laplacian. For graphs drawn from the EPP model, projecting onto the bottom few eigenvectors of the *unnormalized* Laplacian does not work; indeed it is shown by Mihail and Papadimitriou (2002) that high-degree nodes skew the top eigenspace of the adjacency matrix in the direction of the indicator vectors of those vertices. Thus, to cluster such graphs, we must counteract this effect with suitable degree normalization. However, if the minimum degree of any node in the graph is low, then the usual normalized Laplacians have poor concentration properties, and their bottom singular space or eigenspace may not correspond to a subspace where the clusters are well-separated.

A line of previous work (Dasgupta et al., 2004) on clustering graphs generated by this model bypasses this problem by assuming that the generative parameter vector of d_u's is given to the algorithm. Their algorithm and analysis depend critically on the *known parameters* assumption, which does not usually hold in real graph partitioning problems. A second line of work (Coja-Oghlan and Lanka, 2009) addresses poor concentration by eliminating the low degree nodes and clustering the rest of the graph. Their algorithm thus applies to graphs which have a small number of low degree nodes; moreover, they use the adjacency matrix normalized by the product of the degrees, which requires further constraints on the nodes and cluster sizes – see Section 6 for more details.

In this paper, we resolve the issue of poor concentration by introducing the notion of a *degree-corrected* normalized graph Laplacian. For a constant $\tau \geq 0$ and a graph with adjacency matrix $\hat{A}$, the degree-corrected random-walk Laplacian is the matrix $I - (\hat{T} + \tau I)^{-1} \hat{A}$, where $\hat{T}$ is the diagonal matrix of degrees. For $\tau = 0$, the degree-corrected Laplacian reduces to the regular Laplacian. If all the nodes in the graph have high degrees relative to τ, then the bottom k singular subspace of the degree-corrected random walk Laplacian is close to the bottom k singular subspace of the random-walk Laplacian. However, if the graph has a number of low degree nodes, then degree-correction has a regularizing effect on the Laplacian.

Our spectral clustering algorithm projects the nodes in the graph onto the bottom k right singular vectors of the degree-corrected random-walk Laplacian, and clusters the nodes in this subspace. We show guarantees on the performance of this algorithm, demonstrating that it outputs the correct partition under a wide range of parameter values. Unlike Coja-Oghlan and Lanka (2009), our algorithm can find the correct partition even if the cluster sizes are not well-balanced. Our analysis is also tighter than Dasgupta et al. (2004); while our bounds may generally be worse, particularly when the graph has many low degree nodes, under certain conditions, we can show that our performance guarantees are better than those of Dasgupta et al. (2004). Finally, even if the graph has very many low degree nodes, which cannot be reliably clustered because we simply do not have enough adjacent edges available, our algorithm can still use these nodes in the degree-corrected Laplacian to compute a subspace for clustering the high degree nodes reliably.

A key tool in our analysis is a sharp concentration bound on the spectral norm of the degree-corrected random-walk graph Laplacian, which approximately degrades with $\tilde{O}(\frac{1}{\sqrt{\tau}})$. These bounds are then used to show that if the clusters are well-separated, then, after projection onto the bottom k right singular subspace of the degree-corrected random walk Laplacian, nodes from different clusters are well-separated while nodes from the same cluster are close together. A simple thresholding algorithm can then be used to recover the clusters correctly.

Finally, we provide some statistical lower bounds on the performance of any algorithm for finding planted partitions in graphs generated by the EPP model. Our bounds show that when the nodes have uniform degrees, the separation between the clusters required by our algorithm is within a factor of $\tilde{O}(\frac{1}{\sqrt{w_{\min}}})$ of the optimal separation; here $w_{\min}$ is the fraction of nodes that belong to the smallest cluster in the graph.

2. Preliminaries

Planted Partition Model. The planted partition (PP) model is a generative model for random graphs. A graph $G = (V, E)$ generated according to this model has a hidden partition $V_1, \ldots, V_k$ such that $V_1 \cup V_2 \cup \ldots V_k = V$, and $V_i \cap V_j = \emptyset$ for $i \neq j$. If a pair of nodes u and v both lie in some V_i, then, $\Pr[(u, v) \in E] = p$; otherwise $\Pr[(u, v) \in E] = q$. Thus, in the planted partition model, if u and v are two nodes in the same cluster, then their expected degrees are equal.

In the planted partition problem, we are given a graph G generated by the planted partition model, and our goal is to find the hidden partition $V_1, \ldots, V_k$ with high probability over graphs generated according to this model.

Extended Planted Partition Model. The extended planted partition (EPP) model extends this model to graphs with non-uniform degree distributions. A graph $G = (V, E)$ generated according to this model again has a hidden partition $V_1 \cup \ldots \cup V_k = V$. In addition each node u is associated with a number d_u. If two nodes u and v lie in the same cluster V_i, then, $\Pr[(u, v) \in E] = d_u p d_v$; otherwise $\Pr[(u, v) \in E] = d_u q d_v$.

An extended planted partition model is characterized by parameters $(\mathcal{V}, \boldsymbol{d}, p, q)$, where $\mathcal{V} = \{V_1, \ldots, V_k\}$ is the hidden partition, $\boldsymbol{d}$ is the vector of d_u's and p and q are numbers between 0 and 1. We note that the description of a particular model is not unique; for example, for any constant $c > 0$, the parameters $(\mathcal{V}, \boldsymbol{d}, p, q)$ and $(\mathcal{V}, c\boldsymbol{d}, p/c^2, q/c^2)$ describe the same EPP model.

In the extended planted partition problem, we are given a graph G generated by an extended planted partition model, and our goal is to find the hidden partition $V_1, \ldots, V_k$ with high probability over graphs generated according to this model. Observe that unlike the work of Dasgupta et al. (2004), we do not have access to the d_u vector.

Laplacian. The Laplacian of a graph $G = (V, E)$ is defined as the matrix $\hat{L} = \hat{T} - \hat{A}$, where $\hat{T}$ is the diagonal matrix of degrees and $\hat{A}$ is the adjacency matrix of the graph.

Random-walk Laplacian. The random-walk Laplacian of a graph $G = (V, E)$ is defined as the matrix $\hat{\Delta} = I - \hat{T}^{-1}\hat{A}$, where $\hat{T}$ is the diagonal matrix of degrees and $\hat{A}$ is the adjacency matrix of the graph.

Degree-corrected Random-walk Laplacian. The degree-corrected random-walk Laplacian of a graph $G = (V, E)$ is defined as the matrix: $\hat{\Delta}' = I - (\hat{T} + \tau I)^{-1}\hat{A}$, where $\hat{T}$ is the diagonal matrix of degrees, $\hat{A}$ is the adjacency matrix of the graph, and τ is a constant to be specified later.

Normalized Laplacian. The normalized Laplacian of a graph $G = (V, E)$ is defined as the matrix $\hat{\mathcal{L}} = I - \hat{T}^{-1/2}\hat{A}\hat{T}^{-1/2}$, where $\hat{T}$ is the diagonal matrix of degrees and $\hat{A}$ is the adjacency matrix of the graph.

Degree-corrected Normalized Laplacian. The degree-corrected normalized Laplacian of a graph $G = (V, E)$ is defined as the matrix: $\hat{\mathcal{L}}' = I - (\hat{T} + \tau I)^{-1/2}\hat{A}(\hat{T} + \tau I)^{-1/2}$, where $\hat{T}$ is the diagonal matrix of degrees, $\hat{A}$ is the adjacency matrix of the graph, and τ is a constant to be specified later.

Notation. For a node u in graph G, we use the notation $\deg(u)$ to denote the actual degree of u, and $\mathbb{E}[\deg(u)]$ to denote its expected degree.

We use the notation $A = \mathbb{E}[\hat{A}]$ and $T = \mathbb{E}[\hat{T}]$. In addition, we use $\hat{S}$ to denote the $n \times n$ diagonal matrix $\hat{T} + \tau I$, where τ is a constant to be specified later, and S to denote its expectation $\mathbb{E}[\hat{S}]$. We use the notation $\Delta' = I - S^{-1}A$ and $\mathcal{L}' = I - S^{-1/2}AS^{-1/2}$.

For much of the paper, we work with a subgraph of $G = (V, E)$ induced by some subset P of nodes. In this case, we use the subscript P to denote the relevant quantities for this subgraph. For example, $\hat{A}_P$ denotes the adjacency matrix of the subgraph on P, $\deg_P(u)$ denotes the total number of edges between a node u and the nodes in P, and so on.

For a matrix M, we use the notation $\|M\|$ to denote its spectral or L_2 norm. For the rest of the paper, all expectations are taken over graphs drawn from the extended planted partiton model.

We use the notation $\bar{d}$ to denote the average of the d_u's in the graph: $\bar{d} = \frac{1}{n}\sum_{u \in V} d_u$, and for a cluster C_i, we use the notation $\bar{d}_i$ to denote the average of the d_u's among nodes in cluster C_i: $\bar{d}_i = \frac{1}{|C_i|}\sum_{u \in V} d_u$. For a cluster C_i, we use w_i to denote the fraction of nodes in the graph that belong to cluster C_i. We use: $w_{\min} = \min_i w_i$. Moreover, we use the notation $\mathbf{1}$ to denote the all-ones vector of length n, and $\mathbf{1}_j$ to denote the vector of size n the u-th entry of which is 1 if node u belongs to cluster j and 0 otherwise.

We use the notation D to denote the diagonal matrix $\mathbf{diag}(\boldsymbol{d})$.

3. Algorithm

We provide Algorithm 1, an algorithm for finding a planted partition in a graph $G = (V, E)$ drawn from an EPP model. To facilitate the analysis, we split V randomly into two parts P and Q; Q is then projected on to the bottom k right singular subspace of the degree-corrected random-walk Laplacian computed based on P, and partitioned in this subspace; the nodes in P are partitioned analogously. This procedure preserves independence between the subspace-computation and the graph-partitioning steps, thus making the analysis easier.

Observe that Algorithm 1 outputs separate clusterings for P and Q, instead of computing a clustering of the entire graph. Theorem 3 shows that provided certain conditions hold, with high probability, these are correct clusterings of P and Q respectively. In practice however, we may require a clustering of the entire graph; in this case, we can merge the output clusterings into a combined one by merging pairs of clusters C_i of P and C'_j of Q if C_i and C'_j have the closest centers. A second observation is that Algorithm 1 may output some r clusters, where r is not necessarily equal to k; however, if the conditions in Theorem 3 hold, then with probability $1 - \delta$, $r = k$, and the clusters will be the correct clusterings of P and Q.

One can also consider a variant of Algorithm 1 that uses the degree-corrected normalized Laplacian $I - \hat{S}^{-1/2}\hat{A}\hat{S}^{-1/2}$; however, our calculations show that we can get tighter bounds on the separation requirement between the clusters by using the degree-corrected random-walk Laplacian instead.

```
Input: Graph $G = (V, E)$, an integer $k$.
Output: A partition of $V$.

  1. Split $V$ randomly into two sets $P$ and $Q$, each of size $n/2$.
  2. Compute the degree-corrected random-walk Laplacian on $P$
         $\hat{\Delta}'_P = I - \hat{S}_P^{-1} \hat{A}_P,$
     where for $u \in P$, $\hat{S}_P = (\hat{T}_P + \tau I)$, for some $\tau$ to be determined later.
  3. Compute a singular value decomposition of $\hat{\Delta}'_P$. Compute $\hat{U}_P$, the subspace spanned by
     the bottom $k$ *right* singular vectors of $\hat{\Delta}'_P$.
  4. For each node $u$ in $Q$, let $X_u$ be the row of the adjacency matrix corresponding to $u$
     restricted to the nodes in $P$. Let $Y_u = \mathbf{P}_{\hat{U}}(\frac{X_u}{\deg_P(u)})$, and define:
         $\lambda_u = \frac{9\sqrt{k \ln(6kn/\delta)}}{\sqrt{2}(\deg_P(u) - 8\sqrt{\deg_P(u)\ln(6n/\delta)})}$
     Initially, all $u$ in $Q$ are unlabelled.
  5. While there exists an unlabelled node in $Q$:
       (a) Let $u$ be an unlabelled node in $Q$ that maximizes $\deg_P(u)$. Create a new label $l$, and
           assign label $l$ to node $u$.
       (b) For each unlabelled node $v$ in $Q$, if $\|Y_u - Y_v\| \leq \lambda_u + \lambda_v$, assign $v$ the label $l$.
  6. Let $C_l$ be the set of nodes in $Q$ that are labelled $l$. Output clusters $C_1, C_2, \ldots, C_r$.
  7. Repeat Steps (2)-(6) to cluster the nodes in $P$.
```

Algorithm 1: Extended Planted Partition Algorithm

Observe that one difference between Algorithm 1 and McSherry (2001) is that we use the degree-corrected random-walk Laplacian in Step 2; the degree-correction acts as a *regularization* step for the random-walk Laplacian matrix.

A second difference is that we project the vectors $\frac{X_u}{\deg_P(u)}$ instead of X_u onto the bottom k right subspace of the degree-corrected random walk Laplacian computed based on P. Unlike the planted partition model, in EPP, if u and v are drawn from the same partition V_i, then the vectors $\mathbb{E}[X_u]$ and $\mathbb{E}[X_v]$ are no longer equal; instead we have $\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg(u)]} = \frac{\mathbb{E}[X_v]}{\mathbb{E}[\deg(v)]}$. Thus, to ensure that nodes from the same partition are close together after projection, it is necessary to normalize by the degree before projection.

4. Analysis

We now provide performance guarantees for Algorithm 1. We begin with some basic notation. For a cluster V_i, we define the quantity Z_i as: $Z_i = nq\bar{d} + nw_i(p-q)\bar{d}_i$. Observe that for a node u in Q and cluster V_i: $\mathbb{E}[\deg(u)] = \sum_{v \in V_i} d_u p d_v + \sum_{v \notin V_i} d_u q d_v = d_u Z_i$. We define the vector μ_i as: $\mu_i = qD\mathbf{1} + (p-q)D\mathbf{1}_j$.

We use an additional subscript P for these quantities restricted to a subset P of nodes. The notation $\bar{d}_P$ represents the average d_u for nodes $u \in P$, and $\bar{d}_{i,P}$ is the average d_u for nodes $u \in P \cap V_i$. We also use the notation $Z_{i,P}$ and $\mu_{i,P}$ accordingly: $Z_{i,P} = nq\bar{d}_P + nw_{i,P}(p-q)\bar{d}_{i,P}, \mu_{i,P} = qD_P\mathbf{1} + (p-q)D_P\mathbf{1}_j$.

We first analyze clustering the nodes in Q using a projection onto the Laplacian computed based on the nodes in P. The analysis for the other case is analogous.

Theorem 1 *Let $G = (V, E)$ be a random graph drawn from an EPP model. Suppose V can be split into two parts P and Q such that for all u, $\mathbb{E}[\deg_P(u)] \geq \frac{32}{9}\ln(6n/\delta)$. If u and v are two vertices in Q, and if there exists a τ such that for all pairs of clusters i and j,*

$$\begin{aligned}\left\|\frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}}\right\| \quad > \quad & \frac{6\sqrt{\ln(2n/\delta)}}{Z_{i,P}\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2}\right)^{-1/2} \\ & + \frac{6\sqrt{\ln(2n/\delta)}}{Z_{j,P}\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_j \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2}\right)^{-1/2} \\ & + 2 \cdot \left(\min_{u \in V_i \cap Q} \lambda_u + \min_{v \in V_j \cap Q} \lambda_v\right)\end{aligned}$$

then, with probability $1 - 2\delta$, the following statements hold:

1. *If u and v belong to the same cluster in the EPP model, then Step $5(b)$ of Algorithm 1 run with parameter τ assigns them the same label.*
2. *If u and v belong to different clusters in the EPP model, then Step $5(b)$ of Algorithm 1 run with parameter τ assigns them different labels.*

Statements 1 and 2 of Theorem 1 imply that the clustering output by Algorithm 1 is a correct clustering of Q. The theorem involves a parameter τ; the term $\frac{6\sqrt{\ln(2n/\delta)}}{Z_{i,P}\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg_P(u)]}}$ decreases with increasing τ, while $\left(\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2}\right)^{-1/2}$ increases as τ increases. The right hand side of the condition in Theorem 1 is thus optimized when both terms are balanced. The optimal τ has a complex dependence on the degree distribution of the graph; however, for many graphs, we may expect τ to be close to the average degree of G.

Suppose V contains a number of low degree nodes L with large values of λ_u such that the separation conditions in Theorem 1 are satisfied for $V \setminus L$ but not for L. Observe that we can still apply Step 5 of Algorithm 1 on $P \setminus L$ and $Q \setminus L$ to cluster them; the proof of Theorem 1 can be easily extended to show that this will yield the correct clustering. Furthermore, we can still use the nodes in $L \cap P$ to compute the subspace $\hat{U}$ onto which nodes from $Q \setminus L$ can be projected and vice versa, even if we cannot actually cluster the nodes in L reliably.

Theorem 1, combined with Lemma 4 leads to our main theorems. Suppose that the d_u's are all equal; then we have the following result.

Theorem 2 (Main Theorem, uniform *d*) *Let $G = (V, E)$ be a random graph drawn from an extended planted partition model with all d_u's equal to d. Suppose G satisfies the conditions of Lemma 4, q is a constant, and $1 - w_i - w_j$ is at least a constant for all pairs of vertices i and j. If $\tau = 0$, and if:*

$$(p - q) \geq c \cdot \left(\frac{\sqrt{q \ln(2n/\delta)}}{d w_{\min} \sqrt{n}} + \frac{\sqrt{k \ln(6kn/\delta)}}{d^2 \sqrt{n w_{\min}}} \right)$$

where c is a fixed constant, then, w.p. $\geq 1 - 6\delta$, Algorithm 1 outputs correct clusterings of P and Q.

The lower bound on $p - q$ in Theorem 2 has two terms, the first term corresponding to recovering the correct subspace, and the second term corresponding to distance concentration. Our bound is better than the bound of Dasgupta et al. (2004) by a factor of $\sqrt{k}$; we believe that this is an artifact of our analysis. Observe from Theorem 9 that this bound is worse than the statistical lower bound by a factor of $\frac{1}{\sqrt{w_{\min}}}$.

Theorem 2 is a direct consequence of the following more general result:

Theorem 3 (Main Theorem, general case) *Let $G = (V, E)$ be a random graph drawn from an extended planted partition model which satisfies the conditions in Lemma 4. Then, there exists a constant C such that the following holds. If, for all u, $\mathbb{E}[\deg(u)] \geq \frac{128}{9} \ln(6n/\delta)$, and if for all pairs of clusters V_i and V_j,*

$$\left(\frac{p}{Z_i} - \frac{q}{Z_j} \right)^2 \sum_{u \in V_i} d_u^2 + \left(\frac{p}{Z_j} - \frac{q}{Z_i} \right)^2 \sum_{u \in V_j} d_u^2 \geq$$

$$64 \left(\frac{384 \sqrt{\ln(2n/\delta)}}{Z_i \sqrt{\tau + \min_{u \in V_i} \mathbb{E}[\deg(u)]}} \left(\sum_{u \in V_i} \frac{d_u^2}{(\mathbb{E}[\deg(u)] + \tau)^2} \right)^{-1/2} \right.$$

$$\left. + \frac{384 \sqrt{\ln(2n/\delta)}}{Z_j \sqrt{\tau + \min_{u \in V_j} \mathbb{E}[\deg(u)]}} \left(\sum_{u \in V_j} \frac{d_u^2}{(\mathbb{E}[\deg(u)] + \tau)^2} \right)^{-1/2} + \min_{u \in V_i, v \in V_j} 2(\lambda_u + \lambda_v) \right)^2$$

then, w.p. $\geq 1 - 6\delta$, Algorithm 1 outputs a correct clustering.

4.1. Main Lemmas

The main ingredients in the proofs of our main theorems are the following key lemmas.

Lemma 4 *Let $G = (V, E)$ be a random graph drawn from an extended planted partition model with parameters $(\mathcal{V}, \mathbf{d}, p, q)$ such that for any cluster V_i,*

1. $w_i \geq \frac{8 \ln(4k/\delta)}{n}$.

2. $\sum_{u \in V_i} d_u \geq \frac{8}{3}\sqrt{\ln(4k/\delta)}\sqrt{\sum_{u \in V_i} d_u^2}$.
3. $\sum_{u \in V_i} d_u^2 \geq \frac{8}{3}\sqrt{\ln(4k/\delta)}\sqrt{\sum_{u \in V_i} d_u^4}$.
4. *For any* τ, $\sum_{u \in V_i} \frac{d_u^2}{(\mathbb{E}[\deg(u)]+\tau)^2} \geq \frac{8}{3}\sqrt{\ln(4k/\delta)}\sqrt{\sum_{u \in V_i} \frac{d_u^4}{(\mathbb{E}[\deg(u)]+\tau)^4}}$.

Then, with probability $\geq 1-\delta$ *over the splitting of the nodes in* V *into* P *and* Q*, for all clusters* V_i*,*

1. $w_{i,P} \geq \frac{2\ln(4k/\delta)}{n}$.
2. $\sum_{u \in V_i \cap P} d_u \geq \frac{1}{8}\sum_{u \in V_i} d_u$.
3. $\sum_{u \in V_i \cap P} d_u^2 \geq \frac{1}{8}\sum_{u \in V_i} d_u^2$.
4. *For any* τ, $\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)]+\tau)^2} \geq \frac{1}{8}\sum_{u \in V_i} \frac{d_u^2}{(\mathbb{E}[\deg(u)]+\tau)^2}$.

A similar statement also holds for Q.

Lemma 5 *Let* $G = (V, E)$ *be a random graph drawn from an EPP model. Suppose* V *can be split into two parts* P *and* Q *such that for all* u, $\mathbb{E}[\deg_P(u)] \geq \frac{32}{9}\ln(6n/\delta)$*. If* u *and* v *are two vertices in* Q*, and if for all pairs of clusters* i *and* j*,*

$$\begin{aligned} \left\| \frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}} \right\| \quad > \quad & \frac{6\sqrt{\ln(2n/\delta)}}{Z_{i,P}\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2} \right)^{-1/2} \\ & + \frac{6\sqrt{\ln(2n/\delta)}}{Z_{j,P}\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_j \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2} \right)^{-1/2} \\ & + 2 \cdot \left(\min_{u \in V_i \cap Q} \lambda_u + \min_{v \in V_j \cap Q} \lambda_v \right) \end{aligned}$$

then, with probability $1-2\delta$*, the following statements hold:*

1. *If* u *and* v *belong to the same cluster in the EPP model, then* $\|Y_u - Y_v\| \leq \lambda_u + \lambda_v$.
2. *If* u *and* v *belong to different clusters in the EPP model, then* $\|Y_u - Y_v\| > \lambda_u + \lambda_v$.

The proof of Lemma 5 is in turn based on the following three lemmas.

Lemma 6 (Concentration of $\left\|\hat{\Delta}'_P - \Delta'_P\right\|$) *Let* G *be a graph on* n *vertices drawn from the extended planted partition model with parameters* $(\mathcal{V}, \boldsymbol{d}, p, q)$*. Suppose* $\delta \in (0,1)$ *and for each vertex* u, $\mathbb{E}[\deg_P(u)] \geq 6\ln(2n/\delta)$*. Then, with probability* $\geq 1-\delta$*,*

$$\left\|\hat{\Delta}'_P - \Delta'_P\right\| \leq \frac{6\sqrt{\ln(2n/\delta)}}{\sqrt{\tau + \min_{u \in P}\mathbb{E}[\deg(u)]}}.$$

Lemma 7 (Distance Concentration) *Let u be a node in Q, and let X_u be the subset of the row of the adjacency matrix $\hat{A}$ corresponding to node u restricted to the nodes in P. If $\mathbb{E}[\deg_P(u)] \geq \frac{32}{9}\ln(6n/\delta)$, and if U is any fixed k-dimensional subspace, then, with probability $\geq 1-\delta$, for all u,*

$$\left\|\mathbf{P}_U\left(\frac{X_u}{\deg_P(u)}\right) - \mathbf{P}_U\left(\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]}\right)\right\| < \frac{9\sqrt{k\ln(6kn/\delta)}}{\sqrt{2}\mathbb{E}[\deg_P(u)]}.$$

Lemma 8 (Subspace Concentration) *For all clusters i,*

$$\left\|\mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\mu_i) - \mu_i\right\| \leq 2\left\|\Delta'_P - \hat{\Delta}'_P\right\|\left(\sum_{u\in P\cap V_i}\frac{d_u^2}{(d_uZ_i+\tau)^2}\right)^{-1/2}$$

4.2. Proofs of the Main Theorems

Proof (Of Theorem 3) For a set P of vertices, let E_P denote the event that the consequences of Lemma 4 hold true for P. Under the assumption that the conditions Lemma 4 hold, E_P occurs with probability at least $1-\delta$. For the rest of the proof, it is assumed that E_P occurs. We define $\sigma_{u,v} = \lambda_u + \lambda_v$. Recall that

$$\begin{aligned}\left\|\frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}}\right\|^2 &\geq \sum_{v\in V_i\cap P} d_v^2\left(\frac{p}{Z_{i,P}} - \frac{q}{Z_{j,P}}\right)^2 + \sum_{v\in V_j\cap P} d_v^2\left(\frac{p}{Z_{j,P}} - \frac{q}{Z_{i,P}}\right)^2 \\ &= \sum_{v\in V_i\cap P} d_v^2\left(\frac{pZ_{j,P} - qZ_{i,P}}{Z_{i,P}Z_{j,P}}\right)^2 + \sum_{v\in V_j\cap P} d_v^2\left(\frac{pZ_{i,P} - qZ_{j,P}}{Z_{i,P}Z_{j,P}}\right)^2\end{aligned}$$

Recall, $Z_{i,P} \leq Z_i$ and $Z_{j,P} \leq Z_j$. Conditioned on E_P, from Lemma 4, part 2:

$$pZ_{j,P} - qZ_{i,P} = \sum_{u\in V_j\cap P} d_u(p^2-q^2) + \sum_{u\notin V_i\cup V_j, u\in P} d_u q(p-q) \geq \frac{1}{8}(pZ_j - qZ_i)$$

Again conditioning on E_P, combining the two above inequalities with Lemma 4, part 3, we can write:

$$\left\|\frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}}\right\|^2 \geq \frac{1}{64}\left(\left(\frac{p}{Z_i} - \frac{q}{Z_j}\right)^2\sum_{u\in V_i} d_u^2 + \left(\frac{p}{Z_j} - \frac{q}{Z_i}\right)^2\sum_{u\in V_j} d_u^2\right)$$

If the conditions of the theorem hold, then,

$$\begin{aligned}\left\|\frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}}\right\| \quad &\geq \quad \frac{384\sqrt{\ln(2n/\delta)}}{Z_i \cdot \sqrt{\tau + \min_{u \in V} \mathbb{E}[\deg(u)]}} \cdot \left(\sum_{u \in V_i} \frac{d_u^2}{(\mathbb{E}[\deg(u)] + \tau)^2}\right)^{-1/2} \\ &\quad + \frac{384\sqrt{\ln(2n/\delta)}}{Z_j \cdot \sqrt{\tau + \min_{u \in V} \mathbb{E}[\deg(u)]}} \cdot \left(\sum_{u \in V_j} \frac{d_u^2}{(\mathbb{E}[\deg(u)] + \tau)^2}\right)^{-1/2} + 2\sigma_{u,v} \\ &\geq \quad \frac{6\sqrt{\ln(2n/\delta)}}{Z_{i,P} \cdot \sqrt{\tau + \min_{u \in P} \mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2}\right)^{-1/2} \\ &\quad + \frac{6\sqrt{\ln(2n/\delta)}}{Z_{j,P} \cdot \sqrt{\tau + \min_{u \in P} \mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_j \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2}\right)^{-1/2} + 2\sigma_{u,v}.\end{aligned}$$

Here, the second inequality assumes that E_P occurs, following from Lemma 4, part 4.

Again conditioning on E_P, Lemma 4, part 1, implies that if $\mathbb{E}[\deg(u)] \geq \frac{128}{9}\ln(2n/\delta)$, then, $\mathbb{E}[\deg_P(u)] \geq \frac{32}{9}\ln(2n/\delta)$. Hence, the preconditions of Theorem 1 are satisfied. Let C_Q be the event that Algorithm 1 outputs a correct clustering of Q; conditioned on E_P, Theorem 1 implies that C_Q occurs with probability at least $1 - 2\delta$.

We can define the analogous events E_Q and C_P; note that E_P and E_Q are independent. Lemma 4 implies that E_Q occurs with probability at least $1 - \delta$, and conditioned on E_Q, Theorem 1 occurs with probability at least $1 - 2\delta$. C_P and C_Q are independent, and Algorithm 1 correctly clusters both P and Q if both events occur. The theorem follows. ■

Proof (Of Theorem 1) For any pair of nodes u and v in Q, we define $\sigma_{u,v} = \lambda_u + \lambda_v$. Let E be the event that (a) for all pairs u and v that lie in the same cluster V_i, $\|Y_u - Y_v\| \leq \sigma_{u,v}$ and (b) for all pairs u and v that lie in different clusters, $\|Y_u - Y_v\| > \sigma_{u,v}$. Lemma 5 shows that if the conditions of the theorem hold, then, E happens with probability $\geq 1 - 2\delta$. We assume for the rest of the proof that E happens.

We now show the theorem by induction over the iterations of the while loop in Step 5 of Algorithm 1. The induction hypothesis we maintain is that iteration t of Step 5 correctly identifies a partition V_t, and assigns all nodes in $V_t \cap Q$ (and no other nodes) the same label t. The base case is at the beginning when there are no labelled nodes, and hence the induction hypothesis holds trivially.

Suppose the induction hypothesis holds after iteration t of Step 5. This means that t clusters in the graph, say clusters $V_1, \ldots, V_t$ have been correctly identified. Suppose u^* is the node selected in Step $5(a)$ of the next iteration of Step 5; then u^* cannot belong to $V_1 \cup \ldots \cup V_t$. Without loss of generality, let $u^* \in V_{t+1}$. Conditioned on E, if any unlabelled node v belongs to V_{t+1}, then v is assigned label $t + 1$; if $v \notin V_{t+1}$, then v is left unlabelled. Therefore, the cluster V_{t+1} is also recovered correctly. The theorem follows. ■

5. Lower Bounds

In this section, we show a lower bound required on the separation between clusters in the extended planted partition model for any algorithm to be able to correctly discover the clusters. This is a statistical lower bound, in the sense that it depends on statistical properties of the model, regardless of computational considerations.

Theorem 9 *Let $G = (V, E)$ be a graph generated by the EPP model with $k = 3$ and parameters $(\mathcal{V}, \boldsymbol{d}, p, q)$. If $nw_{\min}$ is the minimum size of any cluster in G, then, in order to correctly determine the cluster assignments of all vertices in G w.p. $\geq 3/4$, we need:*

$$(p - q) \geq \frac{\sqrt{\ln 2}}{2d^2\sqrt{3nw_{\min}}}$$

6. Related Work

Spectral clustering has been widely successful as an empirical tool for exploratory data analysis, but despite some prior theoretical work (Ng et al., 2002; von Luxburg, 2007; Balakrishnan et al., 2011), many theoretical aspects remain ill-understood. In this paper, we provide a theoretical analysis of spectral clustering on graphs drawn from a random-graph model, where the nodes are not constrained to have similar degrees. The random graph model we use is called the extended planted partition model, and is a variant of the popular planted partition model.

The planted partition model has long been used as a benchmark for evaluating graph-partitioning algorithms; early work on partitioning random graphs generated by this model include Condon and Karp (2001), Boppana (1987) and Jerrum and Sorkin (1993). The state-of-the-art on the planted partition problem is due to McSherry (2001); he provides a spectral algorithm to recover the planted partitions by using a projection of the nodes onto the top k eigenspace of the adjacency matrix. However, McSherry's work (McSherry, 2001) and those of his predecessors only address the case when all vertices in the same cluster have the same expected degree, and this method fails to recover the correct partition in graphs generated by the extended planted partition model when the degree distribution is too skewed (Mihail and Papadimitriou, 2002).

The extended planted partition model allows for a different expected degree for each node, and as such can be viewed an extension of $G(\boldsymbol{w})$, the *random graph model with given expected degrees* (Chung and Lu, 2006), to the planted partition setting. This extended model was introduced by Dasgupta et al. (2004), who provided a spectral algorithm for partitioning graphs that it generates, *under the assumption that the parameter vector* $\mathbf{d}$ *that generates the graph is known by the algorithm.* This assumption is critical to the algorithm, as its analysis depends on the concentration of the normalized adjacency matrix $D^{-1/2}\hat{A}D^{-1/2}$. When $D = \mathbf{diag}(\boldsymbol{d})$ is unknown, one must normalize $\hat{A}$ by the actual degrees; however, the concentration of this matrix degrades with the minimum degree of any node in the graph. In this paper, we do not assume access to $\boldsymbol{d}$, and we address the concentration problem by using a degree-corrected version of the Laplacian. Our bounds in general can be worse than those of Dasgupta et al. (2004). However, our work improves on Dasgupta et al. (2004) by using the random-walk version of the Laplacian instead of the normalized adjacency matrix; our calculations show that this yields slightly better bounds for partitioning graphs generated by the extended planted partition model.

Coja-Oghlan and Lanka (2009) provide a spectral algorithm for solving the extended planted partition problem which does not require access to $\boldsymbol{d}$, using the top k eigenspace of $\hat{T}^{-1}\hat{A}\hat{T}^{-1}$, where $\hat{T}$ is the diagonal matrix of degrees; however, it only recovers the correct partition under certain conditions – when the maximum expected degree is of lower order than n, the cluster sizes are well-balanced, and the degree of each vertex is at least a constant fraction of the average degree. In contrast, our algorithm succeeds with more imbalance, and does not require these constraints. Moreover, even if there are many low-degree nodes, we can still use them in the degree-corrected Laplacian to find the correct subspace and help cluster the high-degree nodes, even if they cannot be clustered themselves.

Rohe et al. (2011) consider the problem of partitioning graphs generated by the *stochastic block model*, which is the same as the planted partition model; they show that for graphs drawn from this model, the top k eigenvectors of the normalized Laplacian are *consistent*, in the sense that they converge to a "population" limit as the number of nodes n grows to infinity. They also provide guarantees on the performance of a spectral clustering algorithm based on the normalized Laplacian. Unlike our work, they only consider graphs where the expected degrees of nodes in the same cluster are equal. Choi et al. (2011) studies the consistency properties of the maximum-likelihood solution in this model.

Bshouty and Long (2010) provide a nearly linear-time algorithm for partitioning graphs generated by the planted partition model. Their algorithm has faster running time but requires a larger separation conditions. Karrer and Newman (2011) use a variant of the extended planted partition model to probabilistically model graphs, and provide a local heuristic algorithm for estimating the parameters of this model. Theoretical properties of their algorithm are not studied rigorously.

Spectral clustering of *data* drawn from a mixture model is well-understood theoretically; see, for example, Achiloptas and McSherry (2005); Kannan et al. (2005); Kumar and Kannan (2010). This work specifically deals with spectral clustering in *similarity graphs*. Some of the mathematical techniques used in our analysis are related to the work on learning mixture models; examples include Arora and Kannan (2001); Kannan et al. (2005); Achiloptas and McSherry (2005); Chaudhuri and Rao (2008); Kumar and Kannan (2010); however, the techniques for dealing with the effects of varying degrees are specific to this problem. Finally, spectral clustering of random graphs where nodes are data points drawn from a density has been studied by von Luxburg et al. (2008), who provide results on the consistency the graph Laplacian in this setting.

References

D. Achiloptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 458–469, Bertinoro, Italy, 2005.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Annual Symposium on Theory of Computing*, pages 247–257, Heraklion, Greece, 2001.

S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, 2011.

R. B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 280–285, Los Angeles, California, 1987.

N. H. Bshouty and P. Long. Finding planted partitions in nearly linear time using arrested spectral clustering. In *Proceedings of the 27th International Conference on Machine Learning*, pages 135–142, Haifa, Israel, 2010.

K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 9–20, Helsinki, Finland, 2008.

D. Choi, P. Wolfe, and E. Airoldi. Stochastic blockmodels with growing number of classes. *Biometrika*, 2011.

F. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, RI, 1997.

F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, Boston, Massachusetts, 2006.

F. Chung and M. Radcliffe. On the spectra of general random graphs. *Electronic Journal of Combinatorics*, 18(1), 2011.

A. Coja-Oghlan and A. Lanka. Finding planted partitions in random graphs with general degree distributions. *Journal on Discrete Mathematics*, 23(4):1682–1714, 2009.

A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley, New York, 2006.

A. Dasgupta, J. E. Hopcroft, and F. McSherry. Spectral analysis of random graphs with skewed degree distributions. In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, pages 602–610, Rome, Italy, 2004.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

M. Jerrum and G. B. Sorkin. Simulated annealing for graph bisection. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 94–103, Washington, DC, 1993.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 444–457, Bertinoro, Italy, 2005.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 51st Annual Symposium on Foundations of Computer Science*, pages 299–308, Las Vegas, Nevada, 2010.

F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, pages 529–537, Las Vegas, Nevada, 2001.

M. Mihail and C. H. Papadimitriou. On the eigenvalue power law. In *Proceedings of the 6th Interational Workshop on Randomization and Approximation Techniques*, pages 254–262, Cambridge, Massachusetts, 2002.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14(2):849–856, 2002.

R. I. Oliviera. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. Available at arXiv:0911.0600, 2010.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 2011.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36 (2):555–586, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000640.

Appendix A. Appendix

Lemma 10 *Let $X_1, \dots, X_n$ be independent $0/1$ random variables, and $X = \sum_{i=1}^n \alpha_i X_i$, with all $\alpha_i \in (0, 1)$. Let $\|\alpha\|^2 = \sum_{i=1}^n \alpha_i^2$. With probability at least $1 - 2\delta$,*

$$|X - \mathbb{E}[X]| \leq \sqrt{\frac{\|\alpha\|^2 \ln(1/\delta)}{2}}.$$

Proof The proof follows from the standard Hoeffding bound (see Hoeffding (1963)) for independent random variables Y_i, with each $Y_i \in [a_i, b_i]$ and $Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$:

$$\Pr[|Y - E[Y]| \geq \lambda] \leq 2\exp\left(\frac{-2\lambda^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Note that if the X_i's are independent, then $Y_i = \alpha_i X_i$ are as well, and based on this change of variable, $Y_i \in [0, \alpha_i]$ and $Y = \frac{1}{n}X$. Thus, we can write

$$\Pr[|X - E[X]| \geq \lambda] \leq 2\exp\left(\frac{-2\lambda^2}{\sum_{i=1}^{n}\alpha_i^2}\right).$$

The lemma follows by solving for a λ that makes this probability $\leq \delta$. ■

Lemma 11 *Let $X_1, \ldots, X_n$ be independent $0/1$ random variables, and $X = \sum_{i=1}^{n}\alpha_i X_i$, with all $\alpha_i \in (0, 1)$. Let $\nu = \sum_{i=1}^{n}\alpha_i^2 E[X_i]$. If $\nu \geq \frac{1}{18}\ln(1/\delta)$, then with probability at least $1 - 2\delta$,*

$$|X - \mathbb{E}[X]| \leq \frac{1}{3}\ln(1/\delta) + 2\sqrt{\nu\ln(1/\delta)}.$$

Proof Using standard Chernoff bounds appearing in Chung and Lu (2006), we have the following:

$$\Pr[X \leq \mathbb{E}[X] - \lambda] \quad \leq \quad \exp\left(\frac{-\lambda^2}{2\nu}\right), \tag{1}$$

$$\Pr[X \geq \mathbb{E}[X] + \lambda] \quad \leq \quad \exp\left(\frac{-\lambda^2}{2(\nu + \lambda/3)}\right). \tag{2}$$

Setting (1) $\leq \delta$ requires $\lambda \geq \sqrt{2\nu\ln(1/\delta)}$. For the upper tail (2), we need:

$$\lambda \geq \frac{1}{3}\ln(1/\delta) + \sqrt{\frac{1}{9}(\ln(1/\delta))^2 + 2\nu\ln(1/\delta)}.$$

As long as $\nu \geq \frac{1}{18}\ln(1/\delta)$, choosing $\lambda \geq \frac{1}{3}\ln(1/\delta) + 2\sqrt{\nu\ln(1/\delta)}$ gives lower bounds of δ for each tail; the lemma follows. Note that if all $\alpha_i = 1$, then this is a special case and $\nu = \mathbb{E}[X]$. ■

Lemma 12 *Let $X_1, \ldots, X_n$ be independent $0/1$ random variables, and $X = \sum_{i=1}^{n}\alpha_i X_i$, with all $\alpha_i > 0$. Let $\nu = \sum_{i=1}^{n}\alpha_i^2 E[X_i]$. If $(\mathbb{E}[X])^2 \geq \frac{32\nu}{9}\ln(1/\delta)$, then with probability at least $1 - \delta$,*

$$X \geq \frac{1}{4}\mathbb{E}[X].$$

Proof Follows directly from the standard Chernoff bound (1) appearing in Chung and Lu (2006).
■

A.1. Proof of Theorem 2

Proof (Of Theorem 2) Note that when all the d_u's are equal to d, q is a constant, and $1 - w_i - w_j$ is at least a constant, $Z_i = \Theta(ndq)$ and $\mathbb{E}[\deg_P(u)] = \Theta(nd^2q)$. With $\tau = 0$,

$$\begin{aligned}
\left(\frac{p}{Z_i} - \frac{q}{Z_j}\right)^2 \sum_{u \in V_i} d_u^2 &= \Theta\left(\frac{(p-q)^2 w_i}{nq^2}\right), \\
\frac{384\sqrt{\ln(2n/\delta)}}{Z_i\sqrt{\tau + \min_{u\in V_i}\mathbb{E}[\deg(u)]}} \left(\sum_{u\in V_i} \frac{d_u^2}{(\mathbb{E}[\deg(u)]+\tau)^2}\right)^{-1/2} &= \Theta\left(\frac{\sqrt{\ln(2n/\delta)}}{nd\sqrt{qw_i}}\right), \\
\lambda_u &= \Theta\left(\frac{\sqrt{k\ln(6kn/\delta)}}{nd^2q}\right).
\end{aligned}$$

These asymptotic bounds are symmetric for the terms involving V_j. The theorem follows by applying Theorem 3. ■

A.2. Proofs of the main lemmas

Proof (Of Lemma 6) Our proof uses matrix concentration tools from (Chung and Radcliffe, 2011), (Tropp, 2011) and (Oliviera, 2010).

In particular, following Chung and Radcliffe (2011), we can bound $\left\|\hat{\Delta}'_P - \Delta'_P\right\|$ as:

$$\left\|\hat{\Delta}'_P - \Delta'_P\right\| \leq \left\|S_P^{-1}(\hat{A}_P - A_P)\right\| + \left\|S_P^{-1}\hat{A}_P - \hat{S}_P^{-1}\hat{A}_P\right\| \tag{3}$$

To bound the first term in Equation (3), we observe:

$$S_P^{-1}(\hat{A}_P - A_P) = S_P^{-1/2} \cdot S_P^{-1/2}(\hat{A}_P - A_P)S_P^{-1/2} \cdot S_P^{1/2}$$

We can now apply Lemma 15 to conclude:

$$\left\|S_P^{-1}(\hat{A}_P - A_P)\right\| \leq \left\|S_P^{-1/2}(\hat{A}_P - A_P)S_P^{-1/2}\right\|$$

which from Lemma 18 is at most $\sqrt{\frac{3\ln(2n/\delta)}{\tau + \min_{u\in P}\mathbb{E}[\deg_P(u)]}}$ with probability $\geq 1 - \delta/2$.

Let $\hat{\mathcal{L}}'_P = I - \hat{S}_P^{-1/2}\hat{A}_P\hat{S}_P^{-1/2}$. Then, $\hat{A}_P = \hat{S}_P^{1/2}(I - \hat{\mathcal{L}}'_P)\hat{S}_P^{1/2}$. To bound the second term in Equation (3), we observe:

$$\begin{aligned}
\hat{S}_P^{-1}\hat{A}_P - S_P^{-1}\hat{A}_P &= \hat{S}_P^{-1/2}(I - \hat{\mathcal{L}}'_P)\hat{S}_P^{1/2} - S_P^{-1}\hat{S}_P^{1/2}(I - \hat{\mathcal{L}}'_P)\hat{S}_P^{1/2} \\
&= (\hat{S}_P^{-1/2} - S_P^{-1}\hat{S}_P^{1/2}) \cdot (I - \hat{\mathcal{L}}'_P) \cdot \hat{S}_P^{1/2} \\
&= \hat{S}_P^{-1/2} \cdot (I - \hat{S}_P^{1/2}S_P^{-1}\hat{S}_P^{1/2}) \cdot (I - \hat{\mathcal{L}}'_P) \cdot \hat{S}_P^{1/2}
\end{aligned}$$

Now we can again apply Lemma 15 to conclude that:

$$\left\|\hat{S}_P^{-1}\hat{A}_P - S_P^{-1}\hat{A}_P\right\| \leq \left\|I - \hat{S}_P^{1/2}S_P^{-1}\hat{S}_P^{1/2}\right\| \cdot \left\|I - \hat{\mathcal{L}}'_P\right\|$$

Recall that from Lemma 17, $\left\| I - \hat{\mathcal{L}}'_P \right\| \leq 1$. The theorem follows by combining this fact with Lemma 16 with error bound $\delta/2$, noting that $|P| = n/2$. ■

Proof (Of Lemma 7) We can write:

$$\left\| \mathbf{P}_U \left(\frac{X_u}{\deg_P(u)} \right) - \mathbf{P}_U \left(\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]} \right) \right\| \leq \left\| \mathbf{P}_U \left(\frac{X_u - \mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]} \right) \right\| + \left\| \mathbf{P}_U \left(\frac{X_u}{\deg_P(u)} - \frac{X_u}{\mathbb{E}[\deg_P(u)]} \right) \right\|$$

To bound the first term, we can use Lemma 13 with $X_u = X$ and $\delta' = \delta/6n$. This implies that w.p. $\geq 1 - \delta/3$,

$$\left\| \mathbf{P}_U \left(\frac{X_u - \mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]} \right) \right\| \leq \frac{1}{\mathbb{E}[\deg_P(u)]} \sqrt{\frac{k \ln(6kn/\delta)}{2}}.$$

To bound the second term, we note that:

$$\begin{aligned} \left\| \mathbf{P}_U \left(\frac{X_u}{\deg_P(u)} - \frac{X_u}{\mathbb{E}[\deg_P(u)]} \right) \right\| &\leq \left\| \frac{X_u}{\deg_P(u)} - \frac{X_u}{\mathbb{E}[\deg_P(u)]} \right\| \\ &\leq \|X_u\| \cdot \left(\frac{|\mathbb{E}[\deg_P(u)] - \deg_P(u)|}{\deg_P(u)\mathbb{E}[\deg_P(u)]} \right) \\ &\leq \sqrt{\deg_P(u)} \cdot \left(\frac{|\mathbb{E}[\deg_P(u)] - \deg_P(u)|}{\deg_P(u)\mathbb{E}[\deg_P(u)]} \right) \end{aligned} \tag{4}$$

where the last step follows because as X_u is a $0/1$ vector, $\|X_u\| = \sqrt{\deg_P(u)}$. Now we can use the Chernoff bound (Lemma 11) with $\delta' = \delta/6n$ to conclude that w.p. $\geq 1 - \delta/3$,

$$\begin{aligned} |\deg_P(u) - \mathbb{E}[\deg_P(u)]| &\leq \frac{1}{3} \ln(6n/\delta) + 2\sqrt{\mathbb{E}[\deg_P(u)] \ln(6n/\delta)} \\ &\leq 4\sqrt{\mathbb{E}[\deg_P(u)] \ln(6n/\delta)} \end{aligned} \tag{5}$$

Combining with equation (4) and some algebra gives:

$$\left\| \mathbf{P}_U \left(\frac{X_u}{\deg_P(u)} - \frac{X_u}{\mathbb{E}[\deg_P(u)]} \right) \right\| \leq \frac{4\sqrt{\ln(6n/\delta)}}{\mathbb{E}[\deg_P(u)]\sqrt{\deg_P(u)}}.$$

Finally, we note that using the Chernoff bound in Lemma 12, with probability at least $1 - \delta/3$, for all u, $\deg_P(u) \geq \frac{1}{4}\mathbb{E}[\deg_P(u)]$ as long as $\mathbb{E}[\deg_P(u)] \geq \frac{32}{9} \ln(6n/\delta)$. The lemma follows after more algebra and recognizing that $k \geq 2$, and it all happens with probability $1 - \delta$. ■

Proof (Of Lemma 8) Recall that row u of the matrix $I - L_P = S_P^{-1} A_P$ is:

$$\frac{1}{\mathbb{E}[\deg_P(u)] + \tau} d_u \mu_{i,P} = \frac{d_u}{d_u Z_i + \tau} \mu_{i,P}.$$

From Lemma 14 applied to $(S_P^{-1} A_P)^\top$, there exists a vector β such that: $(S_P^{-1} A_P)^\top \beta = \mu_{i,P}$. Moreover,

$$\|\beta\| = \left(\sum_{u \in P \cap C_i} \frac{d_u^2}{(d_u Z_i + \tau)^2} \right)^{-1/2} \tag{6}$$

Let $\mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}$ be the operator that projects a column vector onto the row-space of $\hat{S}_P^{-1}\hat{A}_P$. We can write:

$$\left\|\mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\mu_{i,P}) - \mu_{i,P}\right\| = \left\|(I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P})(S_P^{-1}A_P)^\top \beta\right\| \leq \left\|(I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P})(S_P^{-1}A_P)^\top\right\| \cdot \|\beta\|$$

Equation (6) provides a bound on $\|\beta\|$. To bound the other term, we can write:

$$\begin{aligned}\left\|(I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P})(S_P^{-1}A_P)^\top\right\| \leq\ & \left\|(I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P})(\hat{S}_P^{-1}\hat{A}_P)^\top\right\| \\ & + \left\|(I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P})((S_P^{-1}A_P)^\top - (\hat{S}_P^{-1}\hat{A}_P)^\top)\right\| \end{aligned} \tag{7}$$

The second term in Equation (7) is at most $\left\|S_P^{-1}A_P - \hat{S}_P^{-1}\hat{A}_P\right\|$; this is because the transformation $I - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}$ cannot increase norms. To bound the first term, observe that $\mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\hat{S}_P^{-1}\hat{A}_P)^\top$ is the best rank k approximation to $\hat{S}_P^{-1}\hat{A}_P$. As $S_P^{-1}A_P$ is rank k,

$$\left\|(\hat{S}_P^{-1}\hat{A}_P)^\top - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\hat{S}_P^{-1}\hat{A}_P)^\top\right\| \leq \left\|\hat{S}_P^{-1}\hat{A}_P - S_P^{-1}A_P\right\|$$

The lemma now follows by simple algebra, recognizing that $\left\|\Delta'_P - \hat{\Delta}'_P\right\| = \left\|\hat{S}_P^{-1}\hat{A}_P - S_P^{-1}A_P\right\|$. ■

Proof (Of Lemma 4) The cluster weight $w_{i,P}$ can be written as a sum independent $0/1$ random variables: $w_{i,P} = \frac{2}{n}\sum_{v\in V} X_v$, where $X_v = 1$ with probability $w_i/2$. Applying the Chernoff bound in Lemma 12 with $\delta = \delta'/4k$ and taking the union bound over all V_i gives the result for $w_{i,P}$ with probability $\geq 1 - \delta/4$.

For $\sum_{u\in V_i\cap P} d_u$, note that it can be written as $\sum_{u\in V_i} d_u X_u$, where $X_u = 1$ with probability $1/2$. We again use the Chernoff bound in Lemma 12 and the union bound with $\delta = \delta'/4k$ to get the result in part 2 with probability $\geq 1 - \delta/4$. Part 3 follows from the same argument, replacing d_u by d_u^2. Part 4 follows from the same argument, replacing d_u by $\frac{d_u^2}{(\mathbb{E}[\deg(u)+\tau])^2}$, and the observation that $\mathbb{E}[\deg_P(u)] \leq \mathbb{E}[\deg(u)]$. Finally, the lemma follows from an union bound over the four parts. ■

Proof (Of Lemma 5) For any pair of nodes u and v in Q, we define $\sigma_{u,v} = \lambda_u + \lambda_v$. Suppose u and v are two vertices in the same cluster V_i of the graph, and suppose that u and v lie in Q. Then, recall that $\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]} = \frac{\mathbb{E}[X_v]}{\mathbb{E}[\deg_P(v)]} = \frac{\mu_{i,P}}{Z_{i,P}}$, and therefore:

$$\|Y_u - Y_v\| \leq \left\|Y_u - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]}\right)\right\| + \left\|Y_v - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mathbb{E}[X_v]}{\mathbb{E}[\deg_P(v)]}\right)\right\|$$

Since the projection $\mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}$ is computed independently of the edges adjacent to u and v, we can apply Lemma 7 and (5) to conclude that $\|Y_u - Y_v\|$ is at most $\sigma_{u,v}$ for all u and v with probability $\geq 1 - \delta$.

Now suppose u lies in cluster V_i and v lies in cluster V_j, and both u and v lie in Q. Then,

$$\begin{aligned}\|Y_u - Y_v\| \geq\ & \left\|\frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}}\right\| - \left\|\frac{\mu_{i,P}}{Z_{i,P}} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mu_{i,P}}{Z_{i,P}}\right)\right\| - \left\|\frac{\mu_{j,P}}{Z_{j,P}} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mu_{j,P}}{Z_{j,P}}\right)\right\| \\ & - \left\|Y_u - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mathbb{E}[X_u]}{\mathbb{E}[\deg_P(u)]}\right)\right\| - \left\|Y_v - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\frac{\mathbb{E}[X_v]}{\mathbb{E}[\deg_P(v)]}\right)\right\|\end{aligned}$$

Again using Lemma 7 and (5),

$$\|Y_u - Y_v\| \geq \left\| \frac{\mu_{i,P}}{Z_{i,P}} - \frac{\mu_{j,P}}{Z_{j,P}} \right\| - \frac{1}{Z_{i,P}} \left\| \mu_{i,P} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\mu_{i,P}\right) \right\| - \frac{1}{Z_{j,P}} \left\| \mu_{j,P} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}\left(\mu_{j,P}\right) \right\| - \sigma_{u,v}.$$

Recall that from Lemmas 6 and 8,

$$\left\| \mu_{i,P} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\mu_{i,P}) \right\| \leq \frac{6\sqrt{\ln(2n/\delta)}}{\sqrt{\tau + \min_{u \in P} \mathbb{E}[\deg_P(u)]}} \cdot \left(\sum_{u \in V_i \cap P} \frac{d_u^2}{(\mathbb{E}[\deg_P(u)] + \tau)^2} \right)^{-1/2}$$

for all i with probability $1 - \delta$; a similar statement is therefore true for $\left\| \mu_{j,P} - \mathbf{P}_{\hat{S}_P^{-1}\hat{A}_P}(\mu_{j,P}) \right\|$. Thus, if the conditions of the theorem hold, then,

$$\|Y_u - Y_v\| > \sigma_{u,v}$$

The lemma follows. ■

A.3. Other lemmas

Lemma 13 *Let U be a k-dimensional subspace and X a $0/1$ random vector. Then, w.p. $\geq 1 - 2\delta'$,*

$$\|\mathbf{P}_U(X - \mathbb{E}[X])\| \leq \sqrt{\frac{k \ln(k/\delta')}{2}}.$$

Proof The projection onto U can be written as:

$$\mathbf{P}_U(X - \mathbb{E}[X]) = \sum_{i=1}^{k} \left(\langle X, v_i \rangle - \langle \mathbb{E}[X], v_i \rangle \right) v_i.$$

Applying the Hoeffding bound (Lemma 10) to $\langle X, v_i \rangle$ with $\delta' = \delta/k$ and taking the union over all i gives that for all i, with probability at least $1 - \delta'$:

$$|\langle X, v_i \rangle - \langle \mathbb{E}[X], v_i \rangle| \leq \sqrt{\frac{\ln(k/\delta')}{2}},$$

noting that $\|\alpha\|^2 = 1$ because v_i is a unit vector. Using the triangle inequality,

$$\|\mathbf{P}_U(X - \mathbb{E}[X])\| \leq \sqrt{\sum_{i=1}^{k} |\langle X, v_i \rangle - \langle \mathbb{E}[X], v_i \rangle|^2},$$

from which the lemma follows. ■

Lemma 14 *Let $\boldsymbol{x} \in \mathbf{R}^n$, $\alpha \in \mathbf{R}^k$ be column vectors, and let M be an $n \times k$ matrix such that the ith column of M is $\alpha_i \boldsymbol{x}$. Then, there exists a column vector $\beta \in \mathbf{R}^k$ such that: $M\beta = \boldsymbol{x}$, and $\|\beta\| = \frac{1}{\|\alpha\|}$.*

Proof Let β be the following vector: $\beta_i = \frac{1}{\|\alpha\|^2}\alpha_i$, for $i \in \{1, \ldots, k\}$. Then, $\|\beta\| = \frac{1}{\|\alpha\|}$. Moreover,

$$M\beta = (\sum_i \alpha_i \beta_i)\boldsymbol{x} = \sum_i \frac{\alpha_i^2}{\|\alpha\|^2}\boldsymbol{x} = \boldsymbol{x}$$

The lemma follows. ■

Lemma 15 *Let M be a symmetric matrix, let H be a diagonal matrix, and let $M' = H^{-1/2}MH^{1/2}$. If x is an eigenvector of M with eigenvalue λ, then:*

1. *$H^{-1/2}x$ is a right eigenvector of M' with eigenvalue λ.*
2. *$\left\|M'^{\top}M'\right\| = \|M\|^2$.*

Proof Let $y = H^{-1/2}x$.

$$M'y = H^{-1/2}MH^{1/2}y = H^{-1/2}Mx = \lambda H^{-1/2}x = \lambda y$$

The first part of the lemma follows. To prove the second part, we observe that y is also an eigenvector of $M'^{\top}M'$ with eigenvalue λ^2. ■

Lemma 16 *Let $G = (V, E)$ be a random graph drawn from an EPP model, let $\hat{S} = (\deg(u) + \tau)$ for $\tau > 0$, and let $S = \mathbb{E}[\hat{S}]$. If for some $\delta' \in (0, 1)$ and all $u \in V$, $\mathbb{E}[\deg(u)] \geq \frac{1}{18}\ln(2n/\delta')$, then, w.p. $\geq 1 - \delta'$,*

$$\left\|I - \hat{S}^{1/2}S^{-1}\hat{S}^{1/2}\right\| \leq \frac{4\sqrt{\ln(2n/\delta')}}{\sqrt{\min_u \mathbb{E}[\deg(u)] + \tau}}$$

Proof We use a Chernoff bound (Lemma 11) on $\deg(u)$ with $\delta = \delta'/2n$ and use the union bound over V; for all u, with probability $\geq 1 - \delta'$,

$$|\deg(u) - \mathbb{E}[\deg(u)]| \leq \frac{1}{3}\ln(2n/\delta') + 2\sqrt{\mathbb{E}[\deg(u)]\ln(2n/\delta')}$$

Therefore,

$$\begin{aligned}
\left\| I - \hat{S}^{1/2} S^{-1} \hat{S}^{1/2} \right\| &= \max_u \left| 1 - \frac{\deg(u) + \tau}{\mathbb{E}[\deg(u)] + \tau} \right| \\
&= \max_u \left| \frac{\mathbb{E}[\deg(u)] - \deg(u)}{\mathbb{E}[\deg(u)] + \tau} \right| \\
&\leq \max_u \frac{\frac{1}{3}\ln(2n/\delta') + 2\sqrt{\mathbb{E}[\deg(u)]\ln(2n/\delta')}}{\mathbb{E}[\deg(u)] + \tau} \\
&\leq \max_u \frac{4\sqrt{\mathbb{E}[\deg(u)]\ln(2n/\delta')}}{\mathbb{E}[\deg(u)] + \tau} \\
&\leq \max_u \frac{4\sqrt{\ln(2n/\delta')}}{\sqrt{\mathbb{E}[\deg(u)] + \tau}},
\end{aligned}$$

from which the lemma follows. ■

Lemma 17 *Let $\hat{\mathcal{L}}'_P$ be the degree-corrected normalized Laplacian; then,*

$$\left\| I - \hat{\mathcal{L}}'_P \right\| \leq 1$$

Proof Recall that:

$$I - \hat{\mathcal{L}}'_P = \hat{S}_P^{-1/2} \hat{A}_P \hat{S}_P^{-1/2} = \hat{S}_P^{-1/2} \hat{T}_P^{1/2} \cdot \hat{T}_P^{-1/2} \hat{A}_P \hat{T}_P^{-1/2} \cdot \hat{T}_P^{1/2} \hat{S}_P^{-1/2}$$

Therefore,

$$\left\| I - \hat{\mathcal{L}}'_P \right\| \leq \left\| \hat{S}_P^{-1/2} \hat{T}_P^{1/2} \right\| \cdot \left\| \hat{T}_P^{-1/2} \hat{A}_P \hat{T}_P^{-1/2} \right\| \cdot \left\| \hat{S}_P^{-1/2} \hat{T}_P^{1/2} \right\|$$

As $I - \hat{T}_P^{-1/2} \hat{A}_P \hat{T}_P^{-1/2}$ is the actual normalized Laplacian of G, its eigenvalues are in $[0, 2]$ (as seen in Chung (1997)), so $\hat{T}_P^{-1/2} \hat{A}_P \hat{T}_P^{-1/2}$ has eigenvalues in $[-1, 1]$ and spectral norm ≤ 1. Moreover, each entry of the diagonal matrix $\hat{S}_P^{-1/2} \hat{T}_P^{1/2}$ is at most 1 and therefore its spectral norm is also ≤ 1. The lemma thus follows. ■

Lemma 18 *If a random graph G is drawn from the EPP model, then, with probability $\geq 1 - \delta$,*

$$\left\| S^{-1/2} (\hat{A} - A) S^{-1/2} \right\| \leq \sqrt{\frac{3\ln(2n/\delta)}{\tau + \min_u \mathbb{E}[\deg(u)]}}$$

Proof We heavily use tools from Chung and Radcliffe (2011). Following Theorem 2 of Chung and Radcliffe (2011), we define E^{uv} to be a matrix in which the (u,v)-th entry and the (v,u)-th entry is 1, and the rest of the entries are 0. Let p_{uv} be the probability that the edge (u,v) exists in the graph, and let Z_{uv} be a $0/1$ random variable which is 1 with probability p_{uv} and 0 with probability $1-p_{uv}$. Then, $\hat{A}=\sum_{u,v} p_{uv}Z_{uv}$.

Let

$$H_{uv} = S^{-1/2}((Z_{uv}-p_{uv})A_{uv})S^{-1/2} = \frac{Z_{uv}-p_{uv}}{\sqrt{S_{uu}S_{vv}}}A_{uv}$$

We observe that for any u and v, $\|H_{uv}\| \leq \frac{1}{\sqrt{S_{uu}S_{vv}}}$. Furthermore, for any $u \neq v$, $\mathbb{E}(H_{uv}^2) = \frac{1}{S_{uu}S_{vv}}(p_{uv}-p_{uv}^2)(A_{uu}+A_{vv})$, and $\mathbb{E}(H_{uu}^2)=0$. Therefore,

$$\begin{aligned}\nu^2 &= \left\|\sum_{u,v}\mathbb{E}(H_{uv}^2)\right\| = \left\|\sum_u\sum_v \frac{p_{uv}-p_{uv}^2}{S_{uu}S_{vv}}A_{uu}\right\| = \max_u\left(\sum_v \frac{p_{uv}-p_{uv}^2}{S_{uu}S_{vv}}\right)\\ &\leq \max_u \sum_v \frac{p_{uv}}{S_{uu}S_{vv}} \leq \frac{1}{\min_v S_{vv}}\sum_v \frac{p_{uv}}{S_{uu}} \leq \frac{1}{\min_v S_{vv}}\end{aligned}$$

Here the second to last step follows because $\sum_v p_{uv} = \mathbb{E}[\deg(u)] \leq S_{uu}$. Similar to the proof of the first part of Theorem 2 of Chung and Radcliffe (2011), the lemma now follows by an application of Theorem 5 of Chung and Radcliffe (2011), with $M=1$, $\nu^2 = \frac{1}{\min_v S_{vv}} = \frac{1}{\tau+\min_u \mathbb{E}[\deg(u)]}$, and $a = \sqrt{\frac{3\ln(2n/\delta)}{\tau+\min_u \mathbb{E}[\deg(u)]}}$. ∎

A.4. Proofs from Section 5

Proof (Of Theorem 9) Fix constants $d>0$, $0<\delta<\frac{1}{6d^2}$, $w_{\min}>0$, and let $p=\frac{1}{2d^2}+\delta$, $q=\frac{1}{2d^2}-\delta$, $\boldsymbol{d}=d\mathbf{1}$. Without loss of generality, let $n=|\mathcal{V}|$ and $w_{\min}$ be such that $nw_{\min}$ and $n(1-2w_{\min})$ are integers, and let A be a fixed subset of $n(1-2w_{\min})$ nodes. For a subset of nodes $S\subset\mathcal{V}\setminus A$ of size $nw_{\min}$, let $\bar{S}=\mathcal{V}\setminus(A\cup S)$. Thus, $\{S,\bar{S},A\}$ is a partition of $\mathcal{V}$, and since A is fixed, the partition is determined by the selection of S. We denote by G_S the EPP with partitions given by a specific S, and let F be the family of all such G_S.

Suppose we are given G generated from some $G_S\in F$, and we have an arbitrary algorithm or estimator $\psi(G)$ for a specific member $i\in F$. Then Fano's inequality Cover and Thomas (2006) gives:

$$\sup_{i\in F}\Pr_i[\psi\neq i] \geq 1-\frac{\beta+\ln 2}{\ln r}, \tag{8}$$

where $\mathbf{KL}(G_S,G_{S'})\leq\beta$ for all $G_S,G_{S'}\in F$, and $r=|F|-1$.

For a specific EPP $G_S\in F$, the probability $\Pr[G]$ of generating G is a product of independent Bernoulli distributions over the edges. Suppose that for a possible edge e, the edge probability in G_S is $\rho(e)$ and in $G_{S'}\neq G_S$ is $\rho'(e)$. From Cover and Thomas (2006), we write

$$\mathbf{KL}(G_S,G_{S'}) = \sum_e \mathbf{KL}(\rho,\rho').$$

For each possible edge e, the KL-divergence is zero if $\rho = \rho'$, and otherwise:

$$\begin{aligned}
\mathbf{KL}(\rho, \rho') &\leq d^2p \ln \frac{d^2p}{d^2q} + (1 - d^2p) \ln \frac{1 - d^2p}{1 - d^2q} \\
&= (1/2 + d^2\delta) \ln \frac{1/2 + d^2\delta}{1/2 - d^2\delta} + (1/2 - d^2\delta) \ln \frac{1/2 - d^2\delta}{1/2 + d^2\delta} \\
&= 2d^2\delta \ln \frac{1/2 + d^2\delta}{1/2 - d^2\delta} \\
&= 2d^2\delta \ln \left(1 + \frac{2d^2\delta}{1/2 - d^2\delta}\right) \leq 2d^2\delta \frac{2d^2\delta}{1/2 - d^2\delta} \\
&\leq 2d^2\delta \frac{2d^2\delta}{1/3} \leq 6(d^2\delta)^2 = \frac{3}{2}(d^2(p-q))^2.
\end{aligned}$$

Let N_e be the number of edges e for which $\rho(e) \neq \rho'(e)$. Because A is fixed, we only need to consider edges within $S \cup \bar{S}$:

$$\mathbf{KL}(G_S, G_{S'}) \leq \frac{3}{2} N_e (d^2(p-q))^2 \leq \frac{3}{2} \binom{2nw_{\min}}{2} (d^2(p-q))^2 \leq 3n^2 w_{\min}^2 (d^2(p-q))^2.$$

To bound $|F|$, we consider the number of ways to split $\mathcal{V} \setminus A$ into S and $\bar{S}$. For clarity, let $x = nw_{\min}$. Then

$$|F| = \frac{1}{2}\binom{2x}{x} \geq \frac{2^{2x-2}}{\sqrt{x}},$$

using Stirling's approximation. Therefore,

$$\log(|F| - 1) \geq (2x - 3)\ln 2 - \frac{1}{2}\ln x \geq \frac{\ln 2}{2} x = \frac{\ln 2}{2} nw_{\min}.$$

Substituting into (8),

$$\sup_{i \in F} \Pr_i[\psi \neq i] \geq 1 - \frac{3n^2 w_{\min}^2 (d^2(p-q))^2}{\ln(2) nw_{\min}},$$

which means that w.p. $\geq 3/4$, there is no algorithm that can discern between G_S and G'_S if:

$$p - q \geq \frac{\sqrt{\ln 2}}{2d^2\sqrt{3nw_{\min}}}.$$

■