
Proceedings Paper:
Peng, P. orcid.org/0000-0003-2700-5699 and Yoshida, Y. (2020) Average sensitivity of spectral clustering. In: KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), 23-27 Aug 2020, Virtual Event, CA, USA. Association for Computing Machinery (ACM) , pp. 1132-1140. ISBN 9781450379984

https://doi.org/10.1145/3394486.3403166

Average Sensitivity of Spectral Clustering

Pan Peng
Department of Computer Science
University of Sheffield
p.peng@sheffield.ac.uk

Yuichi Yoshida
National Institute of Informatics
JST, PRESTO
yyoshida@nii.ac.jp

ABSTRACT

Spectral clustering is one of the most popular clustering methods for finding clusters in a graph, which has found many applications in data mining. However, the input graph in those applications may have many missing edges due to error in measurement, withholding for a privacy reason, or arbitrariness in data conversion. To make reliable and efficient decisions based on spectral clustering, we assess the stability of spectral clustering against edge perturbations in the input graph using the notion of average sensitivity, which is the expected size of the symmetric difference of the output clusters before and after we randomly remove edges.

We first prove that the average sensitivity of spectral clustering is proportional to λ_2/λ_3^2, where λ_i is the i-th smallest eigenvalue of the (normalized) Laplacian. We also prove an analogous bound for k-way spectral clustering, which partitions the graph into k clusters. Then, we empirically confirm our theoretical bounds by conducting experiments on synthetic and real networks. Our results suggest that spectral clustering is stable against edge perturbations when there is a cluster structure in the input graph.

CCS CONCEPTS

• **Information systems** → **Clustering**; • **General and reference** → *Reliability*.

KEYWORDS

Spectral clustering, Laplacian, average sensitivity

1 INTRODUCTION

Spectral clustering is one of the most popular graph clustering methods, which finds tightly connected vertex sets, or *clusters*, in the input graph using the eigenvectors of the associated matrix called the (normalized) Laplacian. It has been used in many applications such as image segmentation [21], community detection in networks [8], and manifold learning [3]. See [25] for a survey on the theoretical background and practical use of spectral clustering.

In those applications, however, the input graph is often untrustworthy, and the decision based on the result of spectral clustering may be unreliable and inefficient. We provide some examples below.

- A social network is a graph whose vertices correspond to users in a social network service (SNS), and two vertices are connected if the corresponding users have a friendship relation. However, users may not report their friendship relations because they do not actively use the SNS, or they want to keep the relationship private.
- A sensor network is a graph whose vertices correspond to sensors allocated in some space, and two vertices are connected if the corresponding sensors can communicate with each other. The obtained sensor network may be untrustworthy because some sensors might be unable to communicate due to power shortage or obstacles temporarily put between them.
- In manifold learning, given a set of vectors, we construct a graph whose vertices correspond to the vectors, and we connect two vertices if the corresponding vectors have a distance below a certain threshold. The choice of the threshold is arbitrary, and the obtained graph can vary with the threshold.

If the output clusters are sensitive to edge perturbations, we may make a wrong decision or incur some cost to cancel or update the decision. Hence, to make a reliable and efficient decision using spectral clustering under these situations, the output clusters should be stable against edge perturbations.

One might be tempted to measure the size of the symmetric difference of the output clustering before and after *adversarial* edge perturbations. However, spectral clustering is sensitive to adversarial edge perturbations. For example, suppose that we have a connected graph G with two "significant" clusters $S, \overline{S}$, i.e., the subgraphs induced by S and $\overline{S}$ are well-connected inside while there are few edges between S and $\overline{S}$. It is known that the spectral clustering will output a set, which, together with its complement, corresponds to a bipartition that is close to the partition $\{S, \overline{S}\}$ [15]. However, deleting all the edges incident to any vertex v will result in a new graph G' on which the spectral clustering will output v, or equivalently, the partition $\{\{v\}, \overline{\{v\}}\}$. That is, the output clustering in the original graph is very different from the one in the perturbed graph.

In general, spectral clustering is sensitive to the noisy "dangling sets", which are connected to the core of the graph by only one edge [26]. This suggests that the above way of measuring the stability of spectral clustering might be too pessimistic.

1.1 Our contributions

In this work, we initiate a systematic study of the stability of spectral clustering against edge perturbations, using the notion of average sensitivity [24], which is the expected size of the symmetric difference of the output clusters before and after we *randomly* remove a few edges. Using average sensitivity is more appropriate in many applications, in which the aforementioned adversarial edge perturbations rarely occur. Furthermore, if we can show that the average sensitivity is at most β, then by Markov's inequality, for the 99% of possible edge perturbations, the symmetric difference size of the output cluster is bounded by 100β, which further motivates the use of average sensitivity.

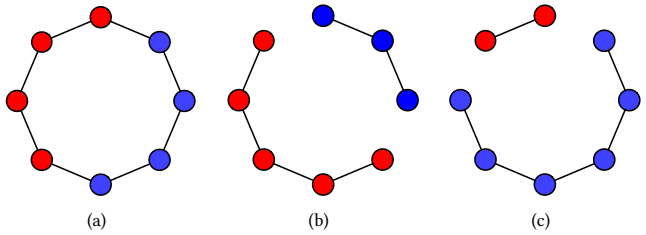


Figure 1: Cycle graph with 8 vertices and the result of 2-way clustering. The output clusters are drastically different between (a) the original graph and (b), (c) the graphs with some edges removed.

We first consider the simplest case of partitioning a graph into two clusters: the algorithm computes the eigenvector corresponding to the second smallest eigenvalue of Laplacian and then outputs a set according to a sweep process over the eigenvector. For both unnormalized and normalized Laplacians, we show that if the input graph has a "significant" cluster structure, then the average sensitivity of spectral clustering is proportional to λ_2/λ_3^2, where λ_i is the i-th smallest eigenvalue of the corresponding Laplacian.

This result is intuitively true because if λ_2/λ_3^2 is small, then by higher-order Cheeger's inequality [15], the graph can be partitioned into two intra-dense clusters with a few edges between them. That is, the cluster structure is significant. Hence, we are likely to output the same cluster after we randomly remove a few edges.

In contrast, a typical graph with a high λ_2/λ_3^2 is an n-cycle (Figure 1). We can observe that the spectral clustering is not stable on this graph because we get drastically different connected components depending on the choice of removed edges.

Next, we consider k-way spectral clustering: the algorithm computes the first k eigenvectors of Laplacian, and then outputs a clustering by invoking the k-means algorithm on the embedding corresponding to the eigenvectors. We consider the spectral clustering algorithm that exploits the normalized Laplacian, which was popularized by Shi and Malik [21]. We show that the average sensitivity of the algorithm is proportional to $\sqrt{\lambda_k}/\lambda_{k+1}$, which again matches the intuition.

Finally, using synthetic and real networks, we empirically confirm that the average sensitivity of spectral clustering correlates well with λ_2/λ_3^2 or $\sqrt{\lambda_k}/\lambda_k$, and that it grows linearly in the edge removal probability, as our theoretical results indicate.

Our theoretical and empirical findings can be summarized as follows.

> We can reliably use spectral clustering: It is stable against random edge perturbations if there is a significant cluster structure in a graph, and it is irrelevant otherwise.

1.2 Organization

We discuss related work in Section 2, and we introduce notions that will be used throughout this paper in Section 3. Then, we study the average sensitivity of 2-way spectral clustering with unnormalized and normalized Laplacians in Sections 4 and 5, respectively. We discuss the average sensitivity of k-way spectral clustering in Section 6. We provide our empirical results in Section 7 and conclude our work in Section 8.

2 RELATED WORK

Huang et al. [12] considered manifold learning, in which we construct a graph G on n vertices from a set of n vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ by connecting the i-th vertex and the j-th vertex by an edge with weight determined by the distance $\|\boldsymbol{v}_i - \boldsymbol{v}_j\|$, and then apply spectral clustering on G. They analyzed how the (weighted) adjacency matrix of G, the normalized Laplacian L of G, and the second eigenvalue/eigenvector of L change when we perturb the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Our work is different in two points. Firstly, we consider random edge perturbations rather than perturbation to the vector data. Hence, we can apply our framework to more general contexts, such as social networks and sensor networks. Secondly, we directly analyze the output clusters rather than the intermediate eigenvalues and eigenvectors.

Bilu and Linial introduced the notion of *stable instances* for clustering problems to model realistic instances with an outstanding solution that is stable under noise [4]. In their setting, for a clustering problem (e.g., sparsest cut) and a parameter $\alpha \geq 1$, a graph with edge or vertex weights is said to be *α-stable* if the optimal solution does not change when we perturb all the edge/vertex weights by a factor of at most α. It has been shown that some clustering problems are solvable in polynomial time on stable instances. Karrer et al. [13] considered the robustness of the community (or cluster) structure of a network by perturbing it as follows: Remove each edge with probability p, and replace it with another edge between a pair of vertices (i, j), chosen at random with an appropriate probability. They used modularity-based methods to find clusters of the original unperturbed graph and the perturbed one, and measured the variation of information between the corresponding two partitions. Gfeller et al. [9] considered the robustness of the cluster structure of weighted graphs where the perturbation is introduced by randomly increasing or decreasing the edge weights by a small factor. In contrast to modeling stable instances for clustering [4] or studying the stability of a partition against random perturbations [9, 13], our study focuses on the stability of the spectral clustering algorithms.

Our work is also related to a line of research on eigenvalue perturbation, which studies how close (or how far) the eigenvalues of a matrix $M + H$ to those of M, where H is "small" in some sense and $M + H$ is viewed as a perturbation of M [22]. The classical theorem due to Weyl (see Theorem 3.6) bounds, for each $i \leq n$, the differences between the i-th eigenvalue of $M + H$ and that of M by the spectral norm of H. Eldridge et al. recently gave some perturbation bounds on the eigenvalues and eigenvectors when the perturbation H is random [6]. There also exist studies on the eigenvalues (and eigenvectors) of $M(\varepsilon)$ that is a matrix function of a parameter ε and is analytic in a small neighborhood of some value ε_0, satisfying $M(\varepsilon_0) = M$ (e.g., [14]). However, we cannot apply such results to our setting, as even for a single edge deletion, we need to consider beyond the small neighborhood.

3 PRELIMINARIES

Let $\mathbf{1} \in \mathbb{R}^n$ be the all-one vector. When all the eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ are real, we denote by $\lambda_i(A)$ the i-th smallest eigenvalue of A. Also, we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the smallest and largest eigenvalues of A, respectively.

We often use the symbols n, m, and Δ to denote the number of vertices, the number of edges, and the maximum degree, respectively, of the graph we are concerned with, which should be clear from the context. For a graph $G = (V, E)$ and a vertex set $S \subseteq V$, $G[S]$ denotes the subgraph of G induced by S. The *volume* $\mathrm{vol}(S)$ of S is the sum of degrees of vertices in S.

3.1 Average Sensitivity

In order to measure the sensitivity of spectral clustering algorithms, which partition the vertex set of a graph into k clusters for $k \geq 2$, we adapt the original definition of average sensitivity [24] as follows.

Let $G = (V, E)$ be a graph. For an edge set $F \subseteq E$, we denote by $G - F$ the graph $(V, E \setminus F)$. For $p \in [0, 1]$, we mean by $F \sim_p E$ that each edge in E is included in F with probability p independently from other edges.

For vertex sets $S, T \subseteq V$, Let $S \triangle T$ denote the symmetric difference between two vertex sets S, T. Let $\mathcal{P} = \{P_1, \ldots, P_k\}$ and $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$ be two k-partitions of V. Then, the *distance* of $\mathcal{P}$ and $\mathcal{Q}$ with respect to vertex size (resp., volume) is defined as follows:

$$d_{\mathrm{size}}(\mathcal{P}, \mathcal{Q}) = \min_{\sigma} \sum_{i=1}^{k} |P_i \triangle Q_{\sigma(i)}|,$$

$$(\text{resp., } d_{\mathrm{vol}}(\mathcal{P}, \mathcal{Q}) = \min_{\sigma} \sum_{i=1}^{k} \mathrm{vol}(P_i \triangle Q_{\sigma(i)})),$$

where σ ranges over all bijections $\sigma\colon \{1, \ldots, k\} \to \{1, \ldots, k\}$. It is easy to see that d_{size} and d_{vol} satisfy the triangle inequality. When $k = 2$ and $\mathcal{P} = \{P, \overline{P}\}$ and $\mathcal{Q} = \{Q, \overline{Q}\}$, we simply write $d_{\mathrm{size}}(P, Q)$ and $d_{\mathrm{vol}}(P, Q)$ instead of $d_{\mathrm{size}}(\mathcal{P}, \mathcal{Q})$ and $d_{\mathrm{vol}}(\mathcal{P}, \mathcal{Q})$.

For an algorithm $\mathcal{A}$ that outputs a k-partition and a real-valued function β on graphs, we say that the *p-average sensitivity* with respect to vertex size (resp., volume) of $\mathcal{A}$ is at most β if

$$\mathop{\mathbf{E}}_{F\sim_p E}[d_{\mathrm{size}}(\mathcal{A}(G), \mathcal{A}(G - F))] \leq \beta(G),$$

$$(\text{resp., } \mathop{\mathbf{E}}_{F\sim_p E}[d_{\mathrm{vol}}(\mathcal{A}(G), \mathcal{A}(G - F))] \leq \beta(G)).$$

We note that this definition is different from the original one [24] in that we remove each edge with probability p independently from others whereas the original one removes k edges without replacement for a given integer parameter k.

3.2 Spectral Graph Theory

3.2.1 Notions from spectral graph theory. Let $G = (V, E)$ be an n-vertex graph. We always assume that the vertices are indexed by integers, i.e., $V = \{1, \ldots, n\}$. For a vertex set S, we denote by $\overline{S}$ the complement set $V \setminus S$.

For two vertex sets $S, T \subseteq V$, $E(S, T)$ denotes the set of edges between S and T, that is, $E(S, T) = \{(i, j) \in E : i \in S, j \in T\}$. The *cut ratio* of $S \subseteq V$ is defined to be $\alpha_G(S) = \frac{|E(S,\overline{S})|}{\min\{|S|,|\overline{S}|\}}$, and the *cut ratio* of G is defined to be $\alpha(G) = \min_{\emptyset\subsetneq S\subsetneq V} \alpha_G(S)$. The *conductance* of a set S is defined to be $\phi_G(S) = \frac{|E(S,\overline{S})|}{\min\{\mathrm{vol}(S),\mathrm{vol}(\overline{S})\}}$, and the *conductance* of G is defined to be $\phi(G) = \min_{\emptyset\subsetneq S\subsetneq V} \phi_G(S)$. For an integer k, let $\rho_G(k)$ be the *k-way expansion* defined as

$$\rho_G(k) := \min_{S_1,\ldots,S_k\colon \text{partition of } V} \max_{1\leq i\leq k} \phi_G(S_i).$$

The *adjacency matrix* $A_G \in \{0, 1\}^{n\times n}$ of G is defined as $(A_G)_{ij} = 1$ if and only if $(i, j) \in E$. The *degree matrix* $D_G \in \mathbb{R}^{n\times n}$ of G is the diagonal matrix with $(D_G)_{ii} = d_i$, where d_i is the degree of the i-th vertex. The *Laplacian* $L_G \in \mathbb{R}^{n\times n}$ of G is defined as $L_G = D_G - A_G$. The *normalized Laplacian* $\mathcal{L}_G \in \mathbb{R}^{n\times n}$ of G is defined as $\mathcal{L}_G = D_G^{-1}L_G = I - D_G^{-1}A_G$.[1] It is well known that all the eigenvalue of L_G and $\mathcal{L}_G$ are nonnegative real numbers. We sometimes call L_G the *unnormalized Laplacian* of G.

We omit subscripts if they are clear from the context.

Algorithm 1: Sweep algorithms

Procedure $\mathrm{sweep}_\alpha(\boldsymbol{v})$
  Reorder the vertices in G in non-decreasing order in terms of $\boldsymbol{v}$, i.e., $v_1 \leq v_2 \leq \cdots \leq v_n$;
  return the set with the minimum *cut ratio* among the sets of the form $\{1, \ldots, i\}$ $(1 \leq i \leq n - 1)$.
Procedure $\mathrm{sweep}_\phi(\boldsymbol{v})$
  Reorder the vertices in G in non-decreasing order in terms of $\boldsymbol{v}$, i.e., $v_1 \leq v_2 \leq \cdots \leq v_n$;
  return the set with the minimum *conductance* among the sets of the form $\{1, \ldots, i\}$ $(1 \leq i \leq n - 1)$.

3.2.2 Cheeger's inequality. We can use unnormalized and normalized Laplacians to find vertex sets with small cut ratio and conductance, respectively. Consider procedures in Algorithm 1, which keep adding vertices in the order determined by the given vector $\boldsymbol{v}$, and then return the set with the best cut ratio and conductance. Then, the following inequality is known.

Lemma 3.1 (Cheeger's inequality [1, 2]). *We have*

$$\frac{\lambda_2}{2} \leq \alpha(G) \leq \sqrt{2\Delta\lambda_2} \quad \text{and} \quad \frac{\nu_2}{2} \leq \phi(G) \leq \sqrt{2\nu_2},$$

where $\lambda_2 = \lambda_2(L_G)$ and $\nu_2 = \lambda_2(\mathcal{L}_G)$. In particular, $\alpha(\mathrm{sweep}_\alpha(\boldsymbol{v}_2)) \leq \sqrt{2\Delta\lambda_2}$ and $\phi(\mathrm{sweep}_\phi(\bar{\boldsymbol{v}}_2)) \leq \sqrt{2\nu_2}$ hold, where $\boldsymbol{v}_2$ and $\bar{\boldsymbol{v}}_2$ are the eigenvectors of L_G and $\mathcal{L}_G$ corresponding to λ_2 and ν_2, respectively.

The following variant of Cheeger's inequality is also known.

Lemma 3.2 (Improved Cheeger's inequality [15]). *For any $k \geq 2$, we have*

$$\alpha(\mathrm{sweep}_\alpha(\boldsymbol{v}_2)) = O\left(\frac{k\lambda_2\sqrt{\Delta}}{\sqrt{\lambda_k}}\right), \text{ and } \phi(\mathrm{sweep}_\phi(\bar{\boldsymbol{v}}_2)) = O\left(\frac{k\nu_2}{\sqrt{\nu_k}}\right),$$

where $\lambda_k = \lambda_k(L_G)$, $\nu_k = \lambda_k(\mathcal{L}_G)$, and $\boldsymbol{v}_2$ and $\bar{\boldsymbol{v}}_2$ are the (right) eigenvectors of L_G and $\mathcal{L}_G$ corresponding to λ_2 and ν_2, respectively.

As $\lambda_k \geq \lambda_2$ and $\nu_k \geq \nu_2$, the bounds in Lemma 3.2 are always stronger than those in Lemma 3.1. We note that the original proof of [16] only handled the normalized case. However, their proof can

[1] In some literature, $\mathcal{L}$ is called the *random-walk Laplacian*.

Algorithm 2: Spectral Clustering

Procedure $\mathcal{USC}_2(G)$
  Compute the second eigenvector $\boldsymbol{v}_2$ of L_G;
  return $\text{sweep}_\alpha(\boldsymbol{v}_2)$.
Procedure $\mathcal{NSC}_2(G)$
  Compute the second eigenvector $\boldsymbol{v}_2$ of $\mathcal{L}$;
  return $\text{sweep}_\phi(\boldsymbol{v}_2)$.

Algorithm 3: Spectral Clustering with k clusters

Procedure $\mathcal{NSC}_{k-\text{means}}(G)$
  Compute the first k eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ of $\mathcal{L}_G$;
  Embedding each vertex $u \in V$ to $f(u) = (\boldsymbol{v}_1(u), \ldots, \boldsymbol{v}_k(u))$;
  return k-means$(f(u_1), \ldots, f(u_n))$.

be easily adapted to the unnormalized case, which we discuss in Appendix A.1.

For k-way expansion, the following higher-order Cheeger inequality is known.

Theorem 3.3 ([18]). *We have*

$$\frac{\lambda_k(\mathcal{L}_G)}{2} \le \rho_G(k) \le O\left(k^2\sqrt{\lambda_k(\mathcal{L}_G)}\right).$$

3.2.3 Spectral Clustering. In this work, we consider the spectral clustering algorithms described in Algorithm 2. In the unnormalized case ($\mathcal{USC}_2$), we first compute the second eigenvector $\boldsymbol{v}_2 \in \mathbb{R}^n$ of the Laplacian of the input graph and then return the set computed by running the sweep_α procedure on $\boldsymbol{v}_2$. In the normalized case ($\mathcal{NSC}_2$), we replace $\boldsymbol{v}_2$ with the second eigenvector of the normalized Laplacian and give it to the sweep_ϕ procedure.

Algorithm 3 is a variant of Algorithm 2 that partitions the graph into k clusters. Here, we first compute the top k eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, and embed each vertex $u \in V$ to a point $(\boldsymbol{v}_1(u), \ldots, \boldsymbol{v}_k(u))$ in $\mathbb{R}^k$. Then, we apply the k-means algorithm [19] to obtain k clusters. This algorithm ($\mathcal{NSC}_{k-\text{means}}$), that makes use of $\mathcal{L}_G = I - D_G^{-1}A_G$, was popularized by Shi and Malik [21]. There are also other versions of spectral clustering based on the k-means algorithm (see [25]). We remark that the one we are analyzing is preferred in practice. For example, in the survey [25], it is said "*In our opinion, there are several arguments which advocate for using normalized rather than unnormalized spectral clustering, and in the normalized case to use the eigenvectors of* L_{rw} (i.e., $\mathcal{L}_G$ as we consider here) *rather than those of* L_{sym} (i.e., $I - D_G^{-1/2}A_GD_G^{-1/2}$)."

The following bound is known for $\mathcal{NSC}_{k\text{-means}}$.

Theorem 3.4 ([20], rephrased). *Let G be a graph with $\frac{\lambda_{k+1}(\mathcal{L}_G)}{\rho_G(k)} = \Omega(k^3)$. Let $\{S_1^*, \ldots, S_k^*\}$ be a k-partition of G achieving $\rho_G(k)$ and let $\{A_1, \ldots, A_k\}$ be the output of $\mathcal{NSC}_{k\text{-means}}$. If the approximation ratio of k-means (in terms of the objective function of k-means) is α, then we have*

$$d_{vol}(\{S_1^*, \ldots, S_k^*\}, \{A_1, \ldots, A_k\}) = O\left(\frac{\alpha k^3 \rho_G(k)}{\lambda_{k+1}(\mathcal{L}_G)} \cdot \text{vol}(G)\right).$$

We remark that in [20], the above theorem was stated in terms of the spectral clustering algorithm that uses the eigenvectors of $L_{\text{sym}} := I - D^{-1/2}AD^{-1/2}$, which turns out to be equivalent to $\mathcal{NSC}_{k\text{-means}}$.

3.3 Stable Instances

We introduce the notion of stable instances, which is another tool we need to analyze average sensitivity of spectral clustering.

For $\varepsilon \in (0,1)$, and two sets $S, T \subseteq V$, we call the corresponding bipartitions $\mathcal{S} = \{S, \overline{S}\}$ and $\mathcal{T} = \{T, \overline{T}\}$ are *ε-close* with respect to size (resp., volume) if

$$d_{\text{size}}(\mathcal{S}, \mathcal{T}) \le \varepsilon n \quad (\text{resp., } d_{\text{vol}}(\mathcal{S}, \mathcal{T}) \le \varepsilon \cdot \text{vol}(G)).$$

We call a graph $G = (V, E)$ *(ρ, ε)-stable* with respect to cut ratio (resp., conductance) if any ρ-approximate solution $S \subseteq V$, that is, $\alpha_G(S) \le \rho \cdot \alpha(G)$ (resp., $\phi_G(S) \le \rho \cdot \phi(G)$), is ε-close to any optimal solution with respect to size (resp., volume). The following is known.

Lemma 3.5 (Corollary 4.17 in [16]). *Let $G = (V, E)$ be a graph. For any $\rho \ge 1$, G is*

$$\left(\rho, \Theta\left(\frac{\rho\lambda_2(L_G)\Delta^{1/2}}{\lambda_3(L_G)^{3/2}}\right)\right)\text{-stable with respect to cut ratio, and}$$

$$\left(\rho, \Theta\left(\frac{\rho\lambda_2(\mathcal{L}_G)}{\lambda_3(\mathcal{L}_G)^{3/2}}\right)\right)\text{-stable with respect to conductance.}$$

Although Kwok et al. [16] showed Lemma 3.5 only for the normalized case, we can easily modify the proof for the unnormalized case, which we provide in Appendix A.2.

3.4 Tools from Matrix Analysis

We will make use of the following results.

Theorem 3.6 (Weyl's inequality). *Let $A, H \in \mathbb{R}^{n\times n}$ be symmetric matrices. Let $\{\lambda_i\}, \{\lambda_i'\}$ be the eigenvalues of A and $A + H$, respectively. Then for any $1 \le i \le n$, we have*

$$\left|\lambda_i - \lambda_i'\right| \le \|H\|,$$

where $\|H\|$ is the spectral norm of H.

Theorem 3.7 (Theorem 5.1.1 in [23]). *Let $X = \sum_{i=1}^T X_i$, where $X_i \in \mathbb{R}^{n\times n}$ $(1 \le i \le T)$ are independent random symmetric matrices. Assume that $0 \le \lambda_{\min}(X_i)$ and $\lambda_{\max}(X_i) \le R$ for any $1 \le i \le T$. Let $\mu_{\max} = \lambda_{\max}(\mathbf{E}[X])$. Then for any $\varepsilon > 0$,*

$$\Pr[\lambda_{\max}(X) \ge (1+\varepsilon)\mu_{\max}] \le n\left(\frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu_{\max}/R}$$

4 SPECTRAL CLUSTERING WITH UNNORMALIZED LAPLACIAN

In this section, we analyze the p-average sensitivity of $\mathcal{USC}_2$ for $p \in [0,1]$. For a graph G, let $\lambda_i(G)$ denote the i-th smallest eigenvalue of the Laplacian associated with a graph G, that is, $\lambda_i(G) = \lambda_i(L_G)$. The goal of this section is to show the following under Assumption 4.2, which we will explain in Section 4.1.

Theorem 4.1. *Let $G = (V, E)$ be a graph and $p \in [0, 1]$. If Assumption 4.2 holds, then the p-average sensitivity of $\mathcal{USC}_2$ is*

$$O\left(\frac{\lambda_2(G)}{\lambda_3(G)^2} \cdot \Delta n + 1\right).$$

We discuss Assumption 4.2 and its plausibility in Section 4.1. Before proving Theorem 4.1, we first discuss the sensitivity of the eigenvalues in Section 4.2. Then, we prove Theorem 4.1 in Section 4.3.

4.1 Assumptions

Given a graph G and a value $p \in [0, p]$, the *reliability* $C(p)$ of G is the probability that if each edge fails with probability p, no connected component of G is disconnected as a result [5]. There exists a fully polynomial-time randomized approximation scheme for $C(p)$ [10]. We will derive a bound on p-average sensitivity under the following assumption. Let $p_{\mathrm{fail}} := O(\max\{\lambda_2(G)/\lambda_3(G)^2, n^{-1}\})$.

Assumption 4.2. We assume the following properties hold.

(i) $\lambda_3(G) \geq \max\{24p\Delta, 48 \log n\}$; (ii) $p_{\mathrm{fail}} \leq 1$;
(iii) $C(p) \geq 1 - p_{\mathrm{fail}}$.

Plausibility of Assumptions. Although the conditions in Assumptions 4.2 are technical, they conform to our intuitions about graphs with low average sensitivity. A graph satisfying those conditions is naturally composed of two vertex disjoint intra-dense subgraphs S^* and $\overline{S^*}$, with no or few crossing edges between them. More generally,

- Assumption 4.2(i) and (ii) imply that λ_2 is small but λ_3 is large, which imply that the graph has at most one outstanding sparse cut by the higher-order Cheeger inequality [15, 18]. It has been discussed that the (normal vectors of the) eigenspaces of Laplacian with such a large eigengap are stable against edge perturbations [25]. To better understand Assumption 4.2(i), let us consider an example. Suppose that G can be partitioned into two clusters S and $\overline{S}$ such that $|S| = \Theta(|\overline{S}|)$, and the induced subgraphs $G[S]$ and $G[\overline{S}]$ have conductance at least $\Omega(1)$ (i.e., the cluster structure of G is significant), and the degree of each vertex in both subgraphs is in $[\Delta/4, \Delta]$. Then it holds that $\lambda_2(G[S]), \lambda_2(G[\overline{S}]) = \Omega(\Delta)$ (see e.g. [11]). This further implies that $\lambda_3(G) = \Omega(\Delta)$ [18], and thus G satisfies the assumption as long as $\Delta = \Omega(\log n)$.
- Assumption 4.2 (iii) corresponds to the intuition that each connected component of the graph remains connected after removing a small set of random edges with high probability. If this is not the case, then intuitively the graph contains many "dangling sets" that are loosely connected to the core of the graph, in which case the algorithm is not stable [26].

Note that if the graph G satisfying Assumption 4.2 is not connected, i.e., $\lambda_2(G) = 0$, then (ii) is trivially satisfied. Then essentially the conditions become that the graph has large $\lambda_3(G)$, and thus has two connected components $S^*, \overline{S^*}$ and p is reasonably small that the corresponding perturbation will not disconnect $G[S^*]$ or $G[\overline{S^*}]$ with high probability.

4.2 Average Sensitivity of Eigenvalues of L_G

The goal of this section is to show the following.

Lemma 4.3. *Let $G = (V, E)$ be a graph and $p \in [0, 1]$, and let $F \sim_p E$. If Assumption 4.2(i) holds, then*

$$\lambda_3(G - F) \geq \frac{\lambda_3(G)}{2}$$

holds with probability at least $1 - n^{-10}$.

We define $E_{ab} \in \{0, 1\}^{n \times n}$ as the matrix such that $(E_{ab})_{cd} = 1$ if and only if $c = a$ and $d = b$. For each edge $e = (i, j) \in E$, we let $E_e = E_{ii} + E_{jj} - E_{ij} - E_{ij}$. For a set $F \subseteq E$ of edges, let $E_F = \sum_{e \in F} E_e$, that is, E_F is the Laplacian matrix of the graph with vertex set V and edge set F. Note that

$$L_{G-F} = D_{G-F} - A_{G-F}$$
$$= D - \sum_{(i,j) \in F} (E_{ii} + E_{jj}) - \left(A - \sum_{(i,j) \in F} (E_{ij} + E_{ji})\right) = L_G - E_F.$$

The following directly follows from Theorem 5 in [6].

Lemma 4.4. *Let $G = (V, E)$ be a graph, $F \subseteq E$, and $1 \leq t \leq n$. Let h be such that $\boldsymbol{x}^\top E_F \boldsymbol{x} \leq h$ for any* unit *vector $\boldsymbol{x} \in \mathbb{R}^n$ in $V' := \mathrm{span}(\boldsymbol{v}_t, \ldots, \boldsymbol{v}_n)$, where $\boldsymbol{v}_i$ is the eigenvector corresponding to $\lambda_i(G)$. Then, we have $\lambda_t(G - F) \geq \lambda_t(G) - h$.*

Next, we prove the following.

Lemma 4.5. *Let $G = (V, E)$ be a graph and $p \in [0, 1]$. Then, with probability $1 - n^{-10}$ over $F \sim_p E$, we have*

$$\boldsymbol{x}^\top E_F \boldsymbol{x} \leq \max\{6p\lambda_n(G), 24 \log n\}.$$

for any unit vector $\boldsymbol{x} \in \mathbb{R}^n$.

Proof. For an edge $e \in E$, let X_e be the indicator random variable of the event that e is included in F. Note that

$$E_F = \sum_{e \in F} E_e = \sum_{e \in E} X_e \cdot E_e.$$

By the fact that $\Pr[X_e = 1] = p$, we have $\mathbf{E}[E_F] = \sum_{e \in E} pE_e = pL_G$. Let $\mu_{\max} := \lambda_{\max}(\mathbf{E}[E_F]) = p\lambda_{\max}(L_G) = p\lambda_n$.

Note that the variables X_e ($e \in E$) are independent and thus E_F is a sum of independent random variables $X_e \cdot E_e$. Further note that $0 \leq \lambda_{\min}(X_e \cdot E_e) \leq \lambda_{\max}(X_e \cdot E_e) \leq 1$. Now by the matrix Chernoff bound (Theorem 3.7), for any $\varepsilon > 0$, we have that

$$\Pr[\lambda_{\max}(E_F) \geq (1 + \varepsilon)\mu_{\max}] \leq n\left(\frac{e^\varepsilon}{(1 + \varepsilon)^{1+\varepsilon}}\right)^{\mu_{\max}}$$

If $\mu_{\max} \leq 4 \log n$, then by setting $\varepsilon = \frac{24 \log n}{\mu_{\max}} - 1 \geq 2e - 1$,

$$\Pr[\lambda_{\max}(E_F) \geq (1 + \varepsilon)\mu_{\max}] \leq n2^{-(1+\varepsilon)\mu_{\max}} \leq \frac{1}{n^{10}}$$

If $\mu_{\max} > 4 \log n$, then by setting $\varepsilon = 5$, we have that

$$\Pr[\lambda_{\max}(E_F) \geq (1 + \varepsilon)\mu_{\max}] \leq n2^{-(1+\varepsilon)\mu_{\max}} \leq \frac{1}{n^{10}}$$

Thus with probability at least $1 - n^{-10}$, for any unit vector $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{x}^\top E_F \boldsymbol{x} \leq \max\{6\mu_{\max}, 24 \log n\} = \max\{6p\lambda_n, 24 \log n\}. \qquad \square$$

Proof of Lemma 4.3. By Lemma 4.5, Lemma 4.4 and the fact that $\lambda_n(G) \leq 2\Delta$, it holds that with probability at least $1 - n^{-10}$, $\lambda_3(G - F) \geq \lambda_3(G) - \max\{12p\Delta, 24 \log n\}$. Then, the inequality in the statement of the lemma directly follows from Assumption 4.2(i). $\square$

4.3 Average Sensitivity of $\mathcal{USC}_2$

In this section, we prove Theorem 4.1.

For a graph $G = (V, E)$, we say a set $S^* \subseteq V$ is *an optimum solution of G with respect to cut ratio* if $\alpha_G(S^*) = \alpha(G)$ and $|S^*| \leq |V|/2$. We first show the following.

Lemma 4.6. *Suppose that Assumption 4.2 holds. Let $F \sim_p E$. Let S^* and S_F^* be optimum solutions of G and $G - F$ with respect to cut ratio, respectively. Then the following holds with probability at least $1 - p_{\text{fail}}$:*

- *if G is not connected, then $S_F^* = S^*$ and $(S_F^*, \overline{S_F^*})$ is the unique cut with cut ratio 0;*
- *otherwise, then $\lambda_2(G - F) > 0$, and*

$$\alpha_{G-F}(S_F^*) \leq \alpha_{G-F}(S^*) \leq \text{APP} \cdot \alpha_{G-F}(S_F^*),$$

where $\text{APP} = O\left(\frac{\lambda_2(G)}{\lambda_2(G-F)}\sqrt{\frac{\Delta}{\lambda_3(G)}}\right)$.

Proof. We first consider the former case. Since $\lambda_3(G) \geq \Omega(\log n)$, S^* and $\overline{S^*}$ induce two connected components of G. By Assumption 4.2(iii), with probability $1 - p_{\text{fail}}$, $(G - F)[S^*]$ and $(G - F)[\overline{S^*}]$ are still connected. Thus $S_F^* = S^*$, which corresponds to the unique cut with cut ratio 0.

Now we consider the latter case. By Assumption 4.2(iii), with probability $1 - p_{\text{fail}}$, the resulting graph $G - F$ is connected, i.e., $\lambda_2(G - F) > 0$.

By definition of cut ratio and Lemma 3.2, it holds that

$$\alpha_{G-F}(S_F^*) \leq \alpha_{G-F}(S^*) \leq \alpha_G(S^*) = O\left(\lambda_2(G)\sqrt{\frac{\Delta}{\lambda_3(G)}}\right).$$

Furthermore by Lemma 3.1, $\lambda_2(G-F)/2 \leq \alpha(G-F) = \alpha_{G-F}(S_F^*)$ and thus we have

$$\alpha_{G-F}(S^*) = O\left(\frac{\lambda_2(G)}{\lambda_2(G-F)}\sqrt{\frac{\Delta}{\lambda_3(G)}} \cdot \alpha_{G-F}(S_F^*)\right). \qquad \square$$

Proof of Theorem 4.1. Let $F \sim_p E$. Let S^* and S_F^* be optimum solutions of G and $G - F$ with respect to cut ratio, respectively. Let S (resp. S_F) be the output of $\mathcal{USC}_2(G)$ (resp. $\mathcal{USC}_2(G - F)$). We further let $\mathcal{S}^* = \{S^*, \overline{S^*}\}$ be the bipartitioning corresponding to S^*. We define $\mathcal{S}_F^*, \mathcal{S}, \mathcal{S}_F$ similarly.

Let $\mathcal{E}$ denote the event that all the statements of Lemma 4.3 and 4.6 hold. Then $\Pr[\mathcal{E}] \geq 1 - n^{-10} - p_{\text{fail}} \geq 1 - 2p_{\text{fail}}$. We first assume that $\mathcal{E}$ holds.

In the case that G is not connected, by Lemma 3.1, the partition $\mathcal{S}$ (resp. $\mathcal{S}_F$) is equivalent to $\mathcal{S}^*$ (resp. $\mathcal{S}_F^*$). Thus, by Lemma 4.6,

$$d_{\text{size}}(\mathcal{USC}_2(G), \mathcal{USC}_2(G - F)) = 0.$$

Now, we assume that G is connected. Let APP be the approximation ratio as specified in Lemma 4.6. Let

$$\varepsilon_1 := \Theta\left(\frac{\Delta^{1/2}}{\lambda_3(G)^{1/2}} \cdot \frac{\lambda_2(G)\Delta^{1/2}}{\lambda_3(G)^{3/2}}\right) = \Theta\left(\frac{\lambda_2(G)\Delta}{\lambda_3(G)^2}\right),$$

$$\varepsilon_2 := \Theta\left(\frac{\Delta^{1/2}}{\lambda_3(G-F)^{1/2}} \cdot \frac{\lambda_2(G-F)\Delta^{1/2}}{\lambda_3(G-F)^{3/2}}\right) = \Theta\left(\frac{\lambda_2(G-F)\Delta}{\lambda_3(G-F)^2}\right),$$

$$\varepsilon_3 := \Theta\left(\text{APP} \cdot \frac{\lambda_2(G-F)\Delta^{1/2}}{\lambda_3(G-F)^{3/2}}\right) = \Theta\left(\frac{\lambda_2(G)\Delta}{\sqrt{\lambda_3(G)\lambda_3(G-F)^3}}\right),$$

$$\varepsilon_\star = \max_{1 \leq i \leq 3}\{\varepsilon_i\} = \Theta\left(\max\left\{\frac{\lambda_2(G)\Delta}{\sqrt{\lambda_3(G)\lambda_3(G-F)^3}}, \frac{\lambda_2(G-F)\Delta}{\lambda_3(G-F)^2}\right\}\right).$$

By Lemmas 3.1 and 3.2, S is an $O(\sqrt{\Delta/\lambda_3(G)})$-approximation of S^*. Thus, we have $d_{\text{size}}(\mathcal{S}, \mathcal{S}^*) \leq \varepsilon_1 n$ by Lemma 3.5. Similarly, S_F is an $O(\sqrt{\Delta/\lambda_3(G-F)})$-approximation of S_F^*, and we have $d_{\text{size}}(\mathcal{S}_F, \mathcal{S}_F^*) \leq \varepsilon_2 n$.

By Lemma 4.6, S^* is an APP-approximation of S_F^*, and hence we have $d_{\text{size}}(\mathcal{S}^*, \mathcal{S}_F^*) < \varepsilon_3 n$. Since $\lambda_2(G - F) \leq \lambda_2(G)$ (by the monotone property of λ_2; see e.g., [7]),we have

$$\begin{aligned} d_{\text{size}}(\mathcal{S}, \mathcal{S}_F) &\leq d_{\text{size}}(\mathcal{S}, \mathcal{S}^*) + d_{\text{size}}(\mathcal{S}^*, \mathcal{S}_F^*) + d_{\text{size}}(\mathcal{S}_F^*, \mathcal{S}_F) \\ &\leq 3\varepsilon_\star n \leq \max\left\{\frac{\lambda_2(G)}{\sqrt{\lambda_3(G)\lambda_3(G-F)^3}}, \frac{\lambda_2(G-F)}{\lambda_3(G-F)^2}\right\} \cdot O(\Delta n) \\ &\leq O\left(\frac{\lambda_2(G)}{\lambda_3(G)^2} \cdot \Delta n\right) \end{aligned}$$

Then, we have

$$\begin{aligned} &\mathop{\mathbf{E}}_{F \sim_p E}[d_{\text{size}}(\mathcal{USC}_2(G), \mathcal{USC}_2(G-F))] = \mathop{\mathbf{E}}_{F \sim_p E}[d_{\text{size}}(\mathcal{S}, \mathcal{S}_F)] \\ &\leq O\left(\frac{\lambda_2(G)}{\lambda_3(G)^2} \cdot \Delta n\right) + n \cdot 2p_{\text{fail}} = O\left(\frac{\lambda_2(G)}{\lambda_3(G)^2} \cdot \Delta n + 1\right), \end{aligned}$$

where in the inequality we used the fact that if $\mathcal{E}$ does not hold, then $d_{\text{size}}(S, S_F) \leq n$. $\square$

5 SPECTRAL CLUSTERING WITH NORMALIZED LAPLACIAN

In this section, we analyze the p-average sensitivity of $\mathcal{NSC}_2$ (with respect to volume) for $p \in [0, 1]$. For a graph G, let $\nu_i(G)$ denote the i-th smallest eigenvalue of the normalized Laplacian associated with a graph G, that is, $\nu_i(G) = \lambda_i(\mathcal{L}_G)$. The goal of this section is to show the following under Assumption 5.2, which we will explain in Section 5.1.

Theorem 5.1. *Let $G = (V, E)$ be a graph and $p \in [0, 1]$. If Assumption 5.2 holds, then the p-average sensitivity of $\mathcal{NSC}_2$ with respect to volume is*

$$O\left(\frac{\nu_2(G)}{\nu_3(G)^2} \cdot \text{vol}(G) + 1\right).$$

We discuss Assumption 5.2 and its plausibility in Section 5.1, and then prove Theorem 5.1 in Section 5.2.

5.1 Assumptions

Let $G = (V, E)$ be a graph with minimum degree τ and maximum degree Δ. Recall that $C(p)$ is the reliability of G given that each edge fails with probability p. Let $p'_{\text{fail}} := O(\max\{\nu_2(G)/\nu_3(G)^2, (2m)^{-1}\})$.

Assumption 5.2. We assume the following properties hold.

(i) $\nu_3(G) \geq \Omega(\tau^{-1} \log n)$; (ii) $p'_{\text{fail}} \leq 1$;
(iii) $p \leq O(\Delta^{-1} \log n)$ and $C(p) \geq 1 - p'_{\text{fail}}$.

Plausibility of Assumptions. The Assumption 5.2(i), (ii) and (iii) can be justified similarly as in Section 4.1. Note that (i) implicitly require that the minimum degree $\tau = \Omega(\log n)$, as $\nu_3(G) \le 2$.

5.2 Average Sensitivity of $\mathcal{NSC}_2$

The following gives a bound on the average sensitivity of eigenvalues of normalized Laplacian. We defer the proof to Appendix B.

Theorem 5.3. *Let $G = (V, E)$ be a graph and $F \sim_p E$. If Assumption 5.2(i) and (iii) hold, then we have $\nu_3(G - F) \ge \nu_3(G)/2$ with probability at least $1 - n^{-7}$.*

Now we give the sketch of the proof of Theorem 5.1.

Proof Sketch of Theorem 5.1. The proof is analogous to that in Section 4.3. Here we mainly sketch the differences.

We will consider the optimum solutions S^* and S_F^* of G and $G - F$ with respect to *conductance*, respectively. Let S (resp. S_F) be the output of $\mathcal{NSC}_2(G)$ (resp. $\mathcal{NSC}_2(G - F)$). We define $\varepsilon_1 = \Theta\left(\frac{\nu_2(G)}{\nu_3(G)^2}\right)$, $\varepsilon_2 = \Theta\left(\frac{\nu_2(G-F)}{\nu_3(G-F)^2}\right)$, $\varepsilon_3 = \Theta\left(\frac{\nu_2(G)}{\sqrt{\nu_3(G)\nu_3(G-F)^3}}\right)$, similarly as in the proof of Theorem 4.1.

By Lemmas 3.1, 3.2, and 3.5, for bipartitions $\mathcal{S} = \{S, \overline{S}\}$ and $\mathcal{S}^* = (S^*, \overline{S^*})$, it holds that $d_{\mathrm{vol}}(\mathcal{S}, \mathcal{S}^*) \le \varepsilon_1 \mathrm{vol}(G)$. For bipartitions $\mathcal{S}_F = \{S_F, \overline{S_F}\}$ and $\mathcal{S}_F^* = \{S_F^*, \overline{S_F^*}\}$, it holds that $d_{\mathrm{vol}}(\mathcal{S}_F, \mathcal{S}_F^*) \le \varepsilon_2 \mathrm{vol}(G - F) \le \varepsilon_2 \mathrm{vol}(G)$.

Analogously to the proof of Lemma 4.6, we can show that S^* is a good approximation of S_F^* in G–F, and that $d_{\mathrm{vol}}(\mathcal{S}^*, \mathcal{S}_F^*) \le \varepsilon_3 \mathrm{vol}(G)$. By bounding the expectation as before, we can obtain the p-average sensitivity of $\mathcal{NSC}_2$, as stated in the theorem. □

6 SPECTRAL CLUSTERING WITH k CLUSTERS

In this section, we consider the p-average sensitivity of $\mathcal{NSC}_{k\text{-means}}$. For a graph G, let $\nu_i(G)$ denote the i-th smallest eigenvalue the normalized Laplacian $\mathcal{L}_G$. We now prove the following.

Theorem 6.1. *Let $G = (V, E)$ be a graph and $p \in [0, 1]$. If Assumption 6.2 holds, then the p-average sensitivity of $\mathcal{NSC}_{k\text{-means}}$ with respect to volume is*

$$O\left(\frac{\alpha k^5 \sqrt{\nu_k(G)}}{\nu_{k+1}(G)} \cdot \mathrm{vol}(G) + 1\right),$$

where α is the approximation ratio of k-means.

6.1 Assumptions

Let $G = (V, E)$ be a graph with minimum degree τ and maximum degree Δ. Let $p''_{\mathrm{fail}} := O(\max\{k^5\sqrt{\nu_k(G)}/\nu_{k+1}(G), (2m)^{-1}\})$.

Assumption 6.2. We assume the following properties hold.

(i) $\nu_{k+1}(G) \ge \Omega(\tau^{-1} \log n)$; (ii) $p''_{\mathrm{fail}} \le 1$;
(iii) $p \le O(\Delta^{-1} \log n)$ and $C(p) \ge 1 - p''_{\mathrm{fail}}$;
(iv) $\nu_{k+1}(G)/\rho_G(k) = \Omega(k^3)$.

The plausibility of the above assumptions can be justified almost the same as in Section 4.1 and 5.1, except that we have one additional condition (iv), which further assumes that the input graph has a significant cluster structure, i.e., it has a k-way partition for which every cluster has low conductance.

6.2 Proof of Theorem 6.1

Similar to the proof of Theorem 5.3, we have the following lemma regarding the perturbation of $\nu_{k+1}(G)$. The only difference is that we use our new assumption on $\nu_{k+1}(G)$.

Lemma 6.3. *Let $G = (V, E)$ be a graph and $F \sim_p E$. If Assumption 6.2(i) and (iii) hold, then we have $\nu_{k+1}(G - F) \ge \frac{\nu_{k+1}(G)}{2}$ with probability at least $1 - n^{-7}$.*

We will make use the following lemma whose proof is provided in Appendix C.

Lemma 6.4. *Let $G = (V, E)$ be a graph. For any $c \ge 1$, G is*

$$\left(c, \Theta\left(\frac{ck^3\rho_G(k)}{\nu_{k+1}(G)}\right)\right)\text{-stable}$$

with respect to conductance.

For a graph G, we say a k-partition $\mathcal{S}^* = \{S_1^*, \ldots, S_k^*\}$ an *optimum solution of G with respect to k-way expansion*, if $\rho_G(\mathcal{S}^*) = \rho_G(k)$. Now we give the sketch of the proof of Theorem 6.1.

Proof Sketch of Theorem 6.1. We will consider the optimum solutions $\mathcal{S}^* = \{S_1^*, \ldots, S_k^*\}$ and $\mathcal{S}_F^* = \{S_{1,F}^*, \ldots, S_{k,F}^*\}$ of G and $G - F$ with respect to k-way expansion, respectively. Let the k-partitions $\mathcal{S} = \{S_1, \ldots, S_k\}$ (resp. $\mathcal{S}_F = \{S_{1,F}, \ldots, S_{k,F}\}$) be the output of $\mathcal{NSC}_{k\text{-means}}(G)$ (resp. $\mathcal{NSC}_{k\text{-means}}(G - F)$).

Let $\varepsilon_1 := \Theta\left(\alpha k^3 \cdot \frac{\rho_G(k)}{\nu_{k+1}(G)}\right)$. By the assumption that $\frac{\nu_{k+1}(G)}{\rho_G(k)} = \Omega(k^3)$, and Theorem 3.4, $d_{\mathrm{vol}}(\mathcal{S}, \mathcal{S}^*) \le \varepsilon_1 \mathrm{vol}(G)$.

Let $\varepsilon_2 := \Theta\left(\alpha k^3 \cdot \frac{\rho_{G-F}(k)}{\nu_{k+1}(G-F)}\right)$. By Lemma 6.3, the assumption $\nu_{k+1}(G)/\rho_G(k) = \Omega(k^3)$, and the fact that $\rho_{G-F}(k) \le \rho_G(k)$ (as $\rho_G(k)$ is a monotone property), we have $\nu_{k+1}(G - F)/\rho_{G-F}(k) = \Omega(k^3)$. By Theorem 3.4, $d_{\mathrm{vol}}\left(\mathcal{S}_F, \mathcal{S}_F^*\right) \le \varepsilon_2 \mathrm{vol}(G)$.

Similarly to the proof of Lemma 4.6, by Assumption (iii) and (iv), we can show that with probability $1 - p_{\mathrm{fail}}$, if G contains k connected components, then $\mathcal{S}_F^* = \mathcal{S}^*$; and otherwise, $\nu_k(G - F) > 0$ and

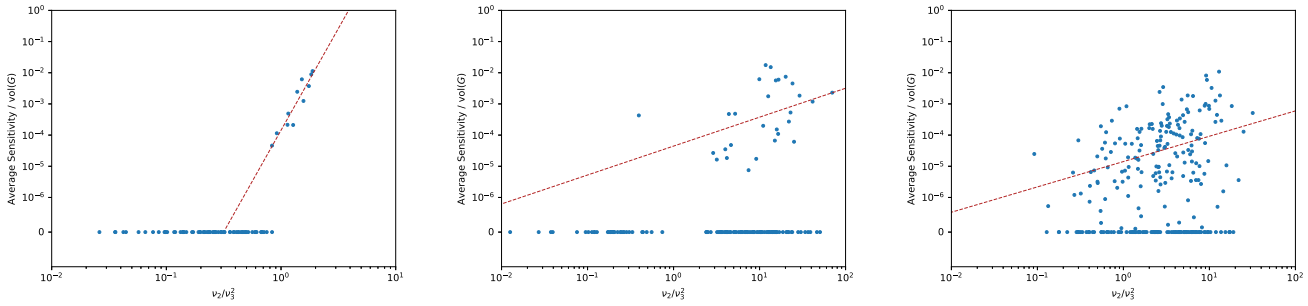
$$\rho_{G-F}(\mathcal{S}_F^*) \le \rho_{G-F}(\mathcal{S}^*) \le \mathrm{APP} \cdot \rho_{G-F}(\mathcal{S}_F^*),$$

where $\mathrm{APP} = \frac{\rho_G(\mathcal{S}^*)}{\rho_{G-F}(\mathcal{S}_F^*)} = \frac{\rho_G(k)}{\rho_{G-F}(k)}$. That is, $\mathcal{S}^*$ is an APP-approximation of $\mathcal{S}_F^*$ in $G - F$.

Thus, we can set $\varepsilon_3 = \Theta\left(k^3 \frac{\rho_G(k)}{\rho_{G-F}(k)} \cdot \frac{\rho_{G-F}(k)}{\nu_{k+1}(G-F)}\right) = \Theta\left(\frac{k^3\rho_G(k)}{\nu_{k+1}(G-F)}\right)$. By Lemma 6.4, $d_{\mathrm{vol}}\left(\mathcal{S}^*, \mathcal{S}_F^*\right) \le \varepsilon_3 \mathrm{vol}(G)$. Thus

$$\begin{aligned}
&d_{\mathrm{vol}}\left(\mathcal{NSC}_{k\text{-means}}(G), \mathcal{NSC}_{k\text{-means}}(G - F)\right) \\
&= d_{\mathrm{vol}}\left(\mathcal{S}, \mathcal{S}_F\right) \le (3 \max_{1 \le i \le 3} \varepsilon_i) \cdot \mathrm{vol}(G) \\
&= O\left(\frac{\alpha k^3 \rho_G(k)}{\nu_{k+1}(G)} \mathrm{vol}(G) + 1\right) = O\left(\frac{\alpha k^5 \sqrt{\nu_k(G)}}{\nu_{k+1}(G)} \cdot \mathrm{vol}(G) + 1\right).
\end{aligned}$$

Finally, by bounding the expectation as before, we can obtain the p-average sensitivity of $\mathcal{NSC}_{k\text{-means}}$ as stated in the theorem. □



(a) SBM2 (b) LFR (c) Twitter

Figure 2: Average sensitivity of $\mathcal{NSC}_{2\text{-means}}$ and ν_2/ν_3^2. Each point represents a graph.

Table 1: Datasets: #, $\overline{n}$, $\overline{m}$, $\overline{\nu_2}$, $\overline{\nu_3}$ are the number of graphs, the average number of vertices, the average number of edges, the average of the second smallest eigenvalue (of the normalized Laplacian), the average of the third smallest eigenvalue, respectively.

Name	#	$\overline{n}$	$\overline{m}$	$\overline{\nu_2}$	$\overline{\nu_3}$
SBM2	80	100.00	1532.04	0.183	0.713
SBM3	80	100.00	1122.86	0.216	0.239
SBM4	80	100.00	918.49	0.233	0.259
LFR	170	100.00	1173.62	0.045	0.165
Twitter	273	131.12	1684.22	0.057	0.169

7 EXPERIMENTS

In this section, we show our experimental results to validate our theoretical results. Here, we focus on spectral clustering with normalized Laplacian ($\mathcal{NSC}_{k\text{-means}}$) because it is advocated for practical use, as we mentioned in Section 3.2.3. We obtained similar results for spectral clustering with unnormalized Laplacian.

As it is computationally hard to calculate the exact value of average sensitivity, we took the average of 1000 trials, where each trial samples a set of edges $F \sim_p E$ and removes F from the graph to compute the symmetric difference size. Also in our plots, we divided the average sensitivity by $\mathrm{vol}(G)$ so that we can compare graphs with different sizes. We set edge removal probability p to be 10^{-3} in all the experiments.

7.1 Datasets

In our experiments, we study five datasets, SBM2, SBM3, SBM4, LFR, and Twitter, which are explained below.

For $k \in \{2, 3, 4\}$, the SBMk dataset is a collection of graphs with k clusters generated with the *stochastic block model*. Specifically, we generate graphs of 100 vertices with k equal-sized clusters by adding an edge for each vertex pair within a cluster with probability p and adding an edge for each vertex pair between different clusters with probability q for each choice of $p \in \{0.3, 0.4, \ldots, 0.9\}$ and $q \in \{0.01, 0.02, \ldots, 0.1\}$.

The LFR dataset is a collection of graphs generated with the Lancichinetti-Fortunato-Radicchi (LFR) benchmark [17]. We run the implementation provided by the authors[2] to generate graphs with 100 vertices, average degree d, the maximum degree 50, and the mixing parameter μ for each choice of $d \in \{4, 5, \ldots, 20\}$ and $\mu \in \{0.01, 0.02, \ldots, 0.1\}$.

The Twitter dataset is a collection of ego-networks in the Twitter network provided at SNAP[3]. As the original dataset was a collection of directed graphs, we discarded directions of edges.

We provide basic information about these datasets in Table 1.

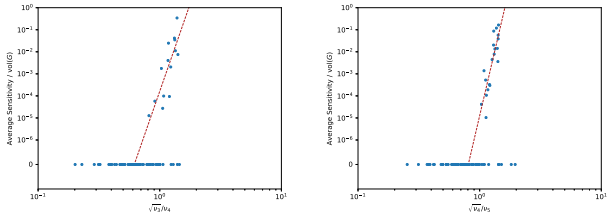
7.2 Results

Average sensitivity of 2-way clustering. Figure 2 shows the relation between p-average sensitivity and ν_2/ν_3^2, where each point represents a graph in the corresponding dataset. The red lines were computed by applying linear regression on graphs with positive average sensitivity. For the SBM2 dataset, we can observe a clear phase transition phenomenon: The average sensitivity dramatically increases when ν_2/ν_3^2 approaches to one. In all the datasets, we can observe that average sensitivity increases as ν_2/ν_3^2 increases. These results empirically confirm the validity of Theorem 5.1.

Average sensitivity of k-way clustering. Figure 3 shows the relation between the average sensitivity of $\mathcal{NSC}_{k-\text{means}}$ and $\sqrt{\nu_k}/\nu_{k+1}$ on the SBMk datasets, which are collections of graphs with k clusters. As with the results for the SBM2 dataset in Figure 2, we can again observe a phase transition phenomenon. These results suggest that the parameter $\sqrt{\nu_k}/\nu_{k+1}$ is critical for the average sensitivity of spectral clustering, as indicated in Theorem 6.1.

Average sensitivity and edge removal probability. Figure 4 shows the relation between average sensitivity grows and edge removal probability p, where a bold line shows the median of average sensitivities of graphs in the corresponding dataset, and the top and bottom of the filled region show the 0.6-quantile and 0.4-quantile, respectively. We can observe that average sensitivity grows almost linear in p. Such a linearity relation has been implicit in our theoretical analysis. For example, under our Assumption 6.2, the proof of Lemma 6.3 (which is similar to the proof of Theorem 5.3) actually gives $\nu_{k+1}(G - F) \geq \nu_{k+1}(G) - 4p$ for

[2] https://www.santofortunato.net/resources
[3] http://snap.stanford.edu/index.html



(a) SBM3 (b) SBM4

Figure 3: Average sensitivity of $\mathcal{NSC}_{k\text{-means}}$ and $\sqrt{\nu_k}/\nu_{k+1}$ on the SBMk dataset. Each point represents a graph.

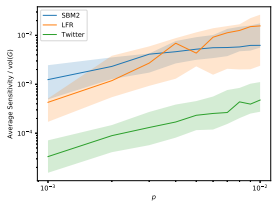


Figure 4: Average sensitivity of $\mathcal{NSC}_{2\text{-means}}$ and edge removal probability p.

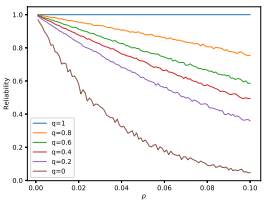


Figure 5: Quantile curves of reliability of graphs in the Twitter dataset.

$p = O(\Delta^{-1}\log n)$ and $\nu_{k+1}(G) \geq \Omega(\tau^{-1}\log n)$. Furthermore, the proof of Theorem 6.1 implies that for small enough p, the average sensitivity of $\mathcal{NSC}_{k\text{-means}}$ is $O\left(\frac{\alpha k^5\sqrt{\nu_k(G)}}{\nu_{k+1}(G-F)}\cdot \mathrm{vol}(G)+1\right) = O\left(\frac{\alpha k^5\sqrt{\nu_k}}{\nu_{k+1}}\cdot \mathrm{vol}(G)\left(1+\frac{5p}{\nu_{k+1}}\right)+1\right)$, which is linear in p.

Reliability. To confirm our assumptions (Assumptions 4.2, 5.2, and 6.2) hold in practice, for each edge failure probability $p \in \{0.001, 0.002, \ldots, 0.1\}$, we computed the reliability of the 273 graphs in the Twitter dataset and calculated the q-quantile of the reliabilities for each $q \in \{0.01, 0.02, \ldots, 1\}$. Figure 5 shows the results. As we can observe, most of the graphs in the Twitter dataset are connected after removing a 10% of edges with high probability, which confirm the plausibility of the assumptions.

8 CONCLUSIONS

To make the decision process more reliable and efficient, we initiate the study of the stability of spectral clustering by using the notion of average sensitivity. We showed that 2-way spectral clustering with both unnormalized and normalized Laplacians has average sensitivity proportional to λ_2/λ_3^2, and k-way spectral clustering with normalized Laplacian has average sensitivity proportional to $\sqrt{\lambda_k}/\lambda_{k+1}$, where λ_i is the i-th smallest eigenvalue of the corresponding Laplacian. We empirically confirmed these theoretical bounds using synthetic and real networks. These results imply that we can reliably use spectral clustering because it is stable against random edge perturbations if there is a significant cluster structure in a graph.

ACKNOWLEDGMENTS

Y. Y. is supported by was supported by JST, PRESTO Grant Number JPMJPR192B, Japan.

REFERENCES

[1] Noga Alon. 1986. Eigenvalues and expanders. *Combinatorica* 6, 2 (1986), 83–96.
[2] Noga Alon and V D Milman. 1985. λ_1, Isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B* 38, 1 (1985), 73–88.
[3] Mikhail Belkin and Partha Niyogi. 2001. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *NIPS*. 585–591.
[4] Yonatan Bilu and Nathan Linial. 2012. Are stable instances easy? *Combinatorics, Probability and Computing* 21, 5 (2012), 643–660.
[5] Charles J Colbourn. 1987. *The combinatorics of network reliability.* Oxford University Press, Inc.
[6] Justin Eldridge, Mikhail Belkin, and Yusu Wang. 2018. Unperturbed: spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory*. 321–358.
[7] Miroslav Fiedler. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* 23, 2 (1973), 298–305.
[8] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3-5 (2010), 75–174.
[9] David Gfeller, Jean-Cédric Chappelier, and Paolo De Los Rios. 2005. Finding instabilities in the community structure of complex networks. *Physical Review E* 72, 5 (2005), 056135.
[10] Heng Guo and Mark Jerrum. 2019. A polynomial-time approximation algorithm for all-terminal network reliability. *SIAM J. Comput.* 48, 3 (2019), 964–978.
[11] Shlomo Hoory, Nathan Linial, and Avi Wigderson. 2006. Expander graphs and their applications. *Bull. Amer. Math. Soc.* 43, 4 (2006), 439–561.
[12] Ling Huang, Donghui Yan, Michael I. Jordan, and Nina Taft. 2008. Spectral Clustering with Perturbed Data. In *NIPS*. 705–712.
[13] Brian Karrer, Elizaveta Levina, and Mark EJ Newman. 2008. Robustness of community structure in networks. *Physical review E* 77, 4 (2008), 046119.
[14] Tosio Kato. 2013. *Perturbation theory for linear operators*. Vol. 132. Springer Science & Business Media.
[15] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. 2013. Improved Cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. In *STOC*. 11–20.
[16] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. 2013. Improved Cheeger's Inequality: Analysis of Spectral Partitioning Algorithms through Higher Order Spectral Gap. *arXiv preprint arXiv:1301.5584* (2013).
[17] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. 2008. Benchmark graphs for testing community detection algorithms. *Physical Review E* 78, 4 (2008).
[18] James R Lee, Shayan Oveis Gharan, and Luca Trevisan. 2014. Multiway spectral partitioning and higher-order cheeger inequalities. *J. ACM* 61, 6 (2014), 37.
[19] J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Vol. I: Statistics, pp. 281–297.
[20] Richard Peng, He Sun, and Luca Zanetti. 2017. Partitioning Well-Clustered Graphs: Spectral Clustering Works! *SIAM J. Comput.* 46, 2 (2017), 710–743.
[21] Jianbo Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905.
[22] G.W. Stewart and J.G. Sun. 1990. *Matrix Perturbation Theory*. ACADEMIC PRESS, INC.
[23] Joel A Tropp et al. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8, 1-2 (2015), 1–230.
[24] Nithin Varma and Yuichi Yoshida. 2019. Average Sensitivity of Graph Algorithms. *CoRR* abs/1904.03248 (2019). arXiv:1904.03248
[25] Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416.
[26] Yilin Zhang and Karl Rohe. 2018. Understanding regularized spectral clustering via graph conductance. In *NeurIPS*. 10631–10640.

A MISSING PROOFS OF SECTION 3

A.1 Proof Sketch of Lemma 3.2

We slightly modify the proof of Lemma 3.2 for the normalized case [16], which is given as Theorem 1.2 in Section 3.1 in [16].

We now start with $\boldsymbol{v}_2$, the second eigenvector of the (unnormalized) Laplacian L, with corresponding second smallest eigenvalue λ_2. By an analogous argument in the proof of Corollary 2.2 of [16], we can find a non-negative function $f : \{1, \ldots, n\} \to \mathbb{R}$ with Rayleigh

quotient $R(f) \le \lambda_2$ and $|\mathrm{supp}(f)| \le n/2$, and $\|f\|_2^2 = 1$, where we identify f with a vector in $\mathbb{R}^n$ and the *Rayleigh quotient* $R\colon \mathbb{R}^n \to \mathbb{R}$ of G is defined as $R(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top L_G \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$.

Then we can find a $(2k+1)$-step function g that well approximates f. By analogous argument in the proof of Lemma 3.1 in [16], we can find such a step function $g\colon \{1,2,\dots,n\} \to \mathbb{R}$ such that $\|f-g\|_2^2 \le \frac{4R(f)}{\lambda_k}$, where λ_k is the k-th smallest eigenvalue of L. (In our case, we need to consider $\|f-g\|_2$, rather than the quantity $\|f-g\|_w$ that takes the degree of each vertex into account as in [16]: e.g., in inequality (3.2) in [16], we do not need the factor $w(v)$. Furthermore, we also need the fact that for any k disjointly supported functions $f_1,\dots,f_k$, we have $\lambda_k \le 2\max_{1\le i\le k} R(f_i)$, whose proof directly follows from the proof of Lemma 2.3 of [16].)

Then, we can also find a function $h\colon \{1,2,\dots,n\} \to \mathbb{R}$ that defines the same sequence of threshold sets as f (and thus we have $\alpha(\mathrm{sweep}_\alpha(h)) = \alpha(\mathrm{sweep}_\alpha(f))$), and satisfies that

$$\frac{\sum_{u\sim v}|h(u)-h(v)|}{\sum_v h(v)} \le 4kR(f) + 4\sqrt{2}k\sqrt{\Delta}\,\|f-g\|_2\sqrt{R(f)}.$$

This follows from an analogous argument for proving the Proposition 3.2 in [16]. (The main difference is that, again, we use the $\|\cdot\|_2$ norm rather than the $\|\cdot\|_w$ norm, and after the third inequality of the proof of Proposition 3.2 on page 16 in [16], we will use the fact that $\sum_{u\sim v}(|f(u)-g(u)|^2 + |(f(v)-g(v))|^2) \le 2\Delta\,\|f-g\|_2^2$.)

Then finally, we use the fact that for any non-negative function h such that $|\mathrm{supp}(h)| \le \frac{n}{2}$, it holds that $\alpha(\mathrm{sweep}_\alpha(h)) \le \frac{\sum_{u\sim v}|h(u)-h(v)|}{\sum_v h(v)}$. (This follows directly from the proof of Lemma 2.4 in [16]). Then, we have $\alpha(\mathrm{sweep}_\alpha(f)) = \alpha(\mathrm{sweep}_\alpha(h))$, which is at most $4k\lambda_2 + 8\sqrt{2}k\sqrt{\Delta}\frac{\lambda_2}{\lambda_k} \le 12\sqrt{2}k\lambda_2\sqrt{\frac{\Delta}{\lambda_k}}$.

A.2 Proof of Lemma 3.5

Consider an arbitrary optimal solution $T \subseteq V$ with $|T| \le n/2$, and let $\alpha = \alpha(T)$. Suppose that there exists a set $S \subseteq V$ with $|S| \le n/2$ and $\alpha(S) \le \rho\alpha$ satisfying that $d_{\mathrm{size}}(S,T) \ge \varepsilon$ for some $0 < \varepsilon \le 1/2$. Let S_1 be either $S \setminus T$ or $T \setminus S$, whichever of a larger size. Let S_2 be either $S \cap T$ or $V \setminus (S \cup T)$, whichever of a larger size. Then by our assumption, we have that $n/2 \ge |S_1| \ge |S \triangle T|/2 \ge \varepsilon n/2$, $|S_2| \ge (n - |S\triangle T|)/2 \ge \varepsilon n/2$ and $|S_2| \le \max\{|S\cap T|, |V\setminus(S\cup T)|\} \le \max\{n/2, (1-\varepsilon)n\} = (1-\varepsilon)n$. Furthermore, for each $i = 1,2$, $|E(S_i, V\setminus S_i)| \le |E(S, V\setminus S)| + |E(T, V\setminus T)| \le \alpha|T| + \rho\alpha|S| \le (1+\rho)\alpha n/2$. Therefore, $\alpha(S_i) \le \frac{(1+\rho)\alpha n/2}{\varepsilon n/2} \le (1+\rho)\alpha/\varepsilon$.

Now let $S_3 = V \setminus (S_1 \cup S_2)$, which is one of the four sets $T, S, V\setminus T, V\setminus S$, and thus $\alpha(S_3) \le \rho\alpha/\varepsilon$. Therefore, $\lambda_3 \le 2\max_i \alpha(S_i) \le \frac{2\rho\alpha}{\varepsilon} \le c_0\frac{\rho\lambda_2}{\varepsilon}\sqrt{\Delta/\lambda_k}$, for some constant $c_0 > 0$. Thus $\varepsilon \le c_0\rho\lambda_2\sqrt{\Delta/\lambda_k^3}$. This further implies that G is $(\rho, \Theta(\rho\lambda_2\sqrt{\Delta/\lambda_3^3}))$-stable.

B PROOF OF THEOREM 5.3

Now we prove Theorem 5.3. Instead of proving that the statement of the theorem holds under Assumption 5.2, we prove it under a weaker assumption.

Let $G = (V,E)$ be a graph with minimum degree τ and $\Gamma := \max_{i\in V}\sum_{j:(j,i)\in E} d_j^{-2}$. Let

$$q := \Theta\left(\max\left\{\sqrt{p\left(\frac{1}{\tau}+\Gamma\right)\log n}, \frac{\log n}{\tau}\right\}\right) \text{ and } p_\diamond := p + q.$$

We need the following assumption:

Assumption B.1. $\nu_3(G) \ge 6p_\diamond$.

Note that any graph G satisfying Assumption 5.2(i) and (iii) also satisfies Assumption B.1. This is true as $\nu_3(G) \ge \Omega(\tau^{-1}\log n)$, $\Gamma \le \frac{\Delta}{\tau^2}$ and $p = O(\Delta^{-1}\log n)$, which gives that

$$p_\diamond = O\left(\max\left\{\sqrt{p\left(\frac{1}{\tau}+\Gamma\right)\log n}, \frac{\log n}{\tau}\right\}\right) \le O\left(\frac{\log n}{\tau}\right),$$

and thus that $\nu_3(G) \ge 6p_\diamond$. Therefore, Theorem 5.3 follows from the following theorem, which we prove in the rest of this section.

Theorem B.2. *Let $G = (V,E)$ be a graph and $F \sim_p E$. If Assumption B.1 holds, then we have $\nu_3(G-F) \ge \nu_3(G)/2$ with probability at least $1 - n^{-7}$.*

We first introduce some definitions. For each $e = (i,j) \in E$, we let $E_{1,e} = E_{ii} + E_{jj}$ and $E_{2,e} = E_{ij} + E_{ji}$. Note that $\sum_{e\in E} E_{1,e} = D$ and $\sum_{e\in E} E_{2,e} = A$. For a set of edges F, let $E_{1,F} = \sum_{e=(i,j)\in F} E_{1,e}$, $E_{2,F} = \sum_{e\in F} E_{2,e}$. Note that $E_F = E_{1,F} - E_{2,F}$ is the unnormalized Laplacian of a graph (V,F). We will make use of the following matrix Chernoff bound.

Theorem B.3 (Corollary 6.1.2 in [23]). *Let $X = \sum_{i=1}^T X_i$, where $X_i \in \mathbb{R}^{n\times n}$ $(1 \le i \le T)$ are independent random matrices. Assume that $\|X_i - \mathbf{E}[X_i]\| \le L$ holds for every $1 \le i \le T$. Let $Y_i = X_i - \mathbf{E}[X_i]$ and $v(X) = \max\left\{\left\|\sum_i \mathbf{E}[Y_i Y_i^\top]\right\|, \left\|\sum_i \mathbf{E}[Y_i^\top Y_i]\right\|\right\}$. Then for any $t > 0$, we have*

$$\Pr[\|X - \mathbf{E}[X]\| \ge t] \le 2n\cdot\exp\left(-\frac{t^2/2}{v(X) + Lt/3}\right).$$

Now we give some claims.

Claim B.4. *With probability at least $1 - n^{-8}$,*

$$\left\|D^{-1}E_{2,F}\right\| \le p_\diamond. \tag{1}$$

Proof. For any $e \in E$, let X_e be the indicator random variable of the event that e is included in F. Note that

$$E_{2,F} = \sum_{e\in F} E_{2,e} = \sum_{e\in E} X_e\cdot E_{2,e} \Longrightarrow D^{-1}E_{2,F} = \sum_{e\in E} X_e\cdot D^{-1}E_{2,e}.$$

Then by $\Pr[X_e = 1] = p$, we have $\mathbf{E}[D^{-1}E_{2,F}] = pD^{-1}\sum_{e\in E} E_{2,e} = pD^{-1}A$.

Now note that the variables X_e $(e \in E)$ are independent and thus $D^{-1}E_{2,F}$ is a sum of independent random variables $S_e := X_e\cdot D^{-1}E_{2,e}$. We further note that $\mathbf{E}[S_e] = pD^{-1}E_{2,e}$.

$$\|S_e - \mathbf{E}[S_e]\| = \left\|(X_e - p)\,D^{-1}E_{2,e}\right\| \le \max\{1-p, p\}\cdot\left\|D^{-1}E_{2,e}\right\| \le \frac{1}{\tau},$$

as we assume that $\min_v \deg(v) \ge \tau$.

Now we note that $E_{2,e}^2 = E_{1,e}$ and thus

$$\begin{aligned}(S_e - \mathbf{E}[S_e])(S_e - \mathbf{E}[S_e])^\top &= (X_e - p)^2 D^{-1}E_{2,e}^2 D^{-1}\\ &= (X_e-p)^2 D^{-1}E_{1,e}D^{-1} = (X_e-p)^2 D^{-2}E_{1,e}.\end{aligned}$$

Thus

$$\mathbf{E}[(S_e - \mathbf{E}[S_e])(S_e - \mathbf{E}[S_e])^\top] = \left(p(1-p)^2 + p^2(1-p)\right)D^{-2}E_{1,e}.$$

Recall that $\sum_e E_{1,e} = D$. We have that

$$\left\|\sum_e \mathbf{E}[(S_e - \mathbf{E}[S_e])(S_e - \mathbf{E}[S_e])^\top]\right\|$$
$$= \left(p(1-p)^2 + p^2(1-p)\right)\left\|\sum_e D^{-2}E_{1,e}\right\| \le p\left\|D^{-1}\right\| \le \frac{p}{\tau}.$$

Similarly, $(S_e - \mathbf{E}[S_e])^\top(S_e - \mathbf{E}[S_e]) = (X_e - p)^2 E_{2,e}D^{-2}E_{2,e}$ holds. Recall that $\Gamma = \max_{i\in V}\sum_{j:(j,i)\in E}\frac{1}{d_j^2}$. Note that $\sum_e E_{2,e}D^{-2}E_{2,e}$ is a diagonal matrix such that each diagonal entry is at most Γ. Then

$$\left\|\sum_e \mathbf{E}[(S_e - \mathbf{E}[S_e])^\top(S_e - \mathbf{E}[S_e])]\right\|$$
$$= \left(p(1-p)^2 + p^2(1-p)\right)\left\|\sum_e E_{2,e}D^{-2}E_{2,e}\right\| \le p\Gamma.$$

By the matrix Chernoff bound (Theorem B.3) with $X_i = S_e$, $X = D^{-1}E_{2,F}$, $\nu(X) = p\max\{1/\tau, \Gamma\}$, $L = 1/\tau$, we have that for any $t > 0$,

$$\Pr\left[\left\|D^{-1}E_{2,F} - pD^{-1}A\right\| \ge t\right] \le 2n\cdot\exp\left(-\frac{t^2/2}{\nu(X) + Lt/3}\right)$$
$$\le 2n\exp\left(-\frac{t^2/2}{p\max\{\frac{1}{\tau},\Gamma\} + \frac{t}{3\tau}}\right).$$

By setting $t = q$, we have that with probability at least $1 - n^{-8}$

$$\left\|D^{-1}E_{2,F} - pD^{-1}A\right\| \le q.$$

Thus,

$$\left\|D^{-1}E_{2,F}\right\| \le p\left\|D^{-1}A\right\| + q \le p + q = p_\diamond. \qquad \square$$

Claim B.5. *With probability at least* $1 - n^{-8}$

$$\left\|D^{-1}E_{1,F}\right\| \le p_\diamond \tag{2}$$

Proof. The proof is similar to the above. Note that

$$D^{-1}E_{1,F} = \sum_{e\in F} D^{-1}E_{1,e} = \sum_{e\in E} X_e\cdot D^{-1}E_{1,e}$$

That is, since the variables X_e ($e \in E$) are independent, $D^{-1}E_{1,F}$ is a sum of independent random variables $S_e := X_e \cdot D^{-1}E_{1,e}$. Note that $\mathbf{E}[S_e] = p\cdot D^{-1}E_{1,e}$. We further note that

$$\mathbf{E}[D^{-1}E_{1,F}] = p\sum_{e\in E} D^{-1}E_{1,e} = p\cdot D^{-1}D = p$$

Then by similar calculations to the proof of the previous claim, we bound $\|S_e - \mathbf{E}[S_e]\| \le \frac{1}{\tau}$ and further by the fact that $E_{1,e}^2 = E_{1,e}$, we have

$$\left\|\sum_e \mathbf{E}[(S_e - \mathbf{E}[S_e])(S_e - \mathbf{E}[S_e])^\top]\right\| \le \frac{p}{\tau}.$$

Now since $E_{1,e}, E_{1,F}, D^{-1}$ are all diagonal matrices, we have

$$\sum_e \mathbf{E}[(S_e - \mathbf{E}[S_e])^\top(S_e - \mathbf{E}[S_e])]$$
$$= \left(p(1-p)^2 + p^2(1-p)\right)\sum_e E_{1,e}D^{-2}E_{1,e}$$
$$= \left(p(1-p)^2 + p^2(1-p)\right)\sum_e E_{1,e}D^{-2}$$

Thus by similar analysis as before, we have

$$\left\|\sum_e \mathbf{E}[(S_e - \mathbf{E}[S_e])^\top(S_e - \mathbf{E}[S_e])]\right\| \le \frac{p}{\tau}.$$

Then the rest follows the same as before. $\square$

Proof of Theorem B.2. We set $p_\diamond$ to be the maximum of the RHS of Ineq. (1) and the RHS of Ineq. (2). Then by the above two claims, we have $\left\|D^{-1}E_{2,F}\right\| \le p_\diamond$ and $\left\|D^{-1}E_{1,F}\right\| \le p_\diamond$ with probability at least $1 - 2\cdot n^{-8}$. We will condition on the above two inequalities in the following.

We have

$$\mathcal{L}_{G-F} = I - D_{G-F}^{-1}A_{G-F} = I - (D - E_{1,F})^{-1}(A - E_{2,F})$$
$$= I - (I - D^{-1}E_{1,F})^{-1}D^{-1}(A - E_{2,F})$$
$$= I - \sum_{i\ge 0}(D^{-1}E_{1,F})^i\cdot D^{-1}(A - E_{2,F}) \quad (\text{by } \left\|D^{-1}E_{1,F}\right\| \le p_\diamond < 1)$$
$$= I - \left(I + \sum_{i\ge 1}(D^{-1}E_{1,F})^i\right)D^{-1}(A - E_{2,F})$$
$$= I - \left(D^{-1} + \sum_{i\ge 1}(D^{-1}E_{1,F})^i D^{-1}\right)(A - E_{2,F})$$
$$= I - D^{-1}A + D^{-1}E_{2,F} - \sum_{i\ge 1}(D^{-1}E_{1,F})^i\cdot D^{-1}(A - E_{2,F})$$
$$= \mathcal{L} + D^{-1}E_{2,F} - \sum_{i\ge 1}(D^{-1}E_{1,F})^i\cdot D^{-1}(A - E_{2,F})$$

Now, we have that

$$\left\|D^{-1}E_{2,F} - \sum_{i\ge 1}(D^{-1}E_{1,F})^i\cdot D^{-1}(A - E_{2,F})\right\|$$
$$\le \left\|D^{-1}E_{2,F}\right\| + \sum_{i\ge 1}\left\|D^{-1}E_{1,F}\right\|^i\left\|D^{-1}(A - E_{2,F})\right\|$$
$$\le p_\diamond + \sum_{i\ge 1}\left\|D^{-1}E_{1,F}\right\|^i \le p_\diamond + \sum_{i\ge 1}p_\diamond^i \le p_\diamond + \frac{3}{2}p_\diamond < 3p_\diamond,$$

where the penultimate inequality follows from the assumption that $p_\diamond \le \frac{1}{6}\nu_3(G) \le \frac{1}{3}$.

Finally, by Weyl's inequality,

$$\nu_3(G - F) \ge \nu_3(G) - \left\|D^{-1}E_{2,F} - \sum_{i\ge 1}(D^{-1}E_{1,F})^i\cdot D^{-1}(A - E_{2,F})\right\|$$
$$> \nu_3(G) - 3p_\diamond \ge \frac{\nu_3(G)}{2},$$

where the last inequality follows from Assumption B.1. $\square$

C MISSING PROOFS OF SECTION 6

Now we sketch the proof of Lemma 6.4.

Proof Sketch of Lemma 6.4. The proof follows by adapting the proof of Lemma 3.5, i.e., Corollary 4.17 in [16], which considers the case $k = 2$. For $k > 2$, we only need to start with an optimum solution $\mathcal{S}^* = \{S_1^*, \ldots, S_k^*\}$ with $\rho_G(\mathcal{S}) = \rho_G(k)$. Then we assume the instance is not (c,ε)-stable, and then construct as in [16] a $(k+1)$-partition $T_1, \ldots, T_{k+1}$ such that $\nu_{k+1}(G) \le 2\max_i \phi(T_i) = O(ck^3\rho_G(k)/\varepsilon)$. Then we conclude that $\varepsilon = O(ck^3\rho_G(k)/\nu_{k+1}(G))$. $\square$