Local Spectral Clustering of Density Upper Level Sets

Alden Green Sivaraman Balakrishnan Ryan Tibshirani

Department of Statistics and Data Science
Carnegie Mellon University

{ajgreen,sbalakri,ryantibs}@andrew.cmu.edu

October 25, 2019

1 Lower bound.

To show a lower bound for density clustering using PPR, we exhibit a hard case: that is, a distribution $\mathbb{P}$ for which PPR is unlikely to recover a density cluster. Let $\mathcal{C}_\sigma^{(0)}$, $\mathcal{C}_\sigma^{(1)}$, and $\mathcal{C}_\sigma^{(2)}$ be rectangles in $\mathbb{R}^2$,

$$\mathcal{C}_\sigma^{(0)} = \left[-\frac{\sigma}{2}, \frac{\sigma}{2}\right] \times \left[-\frac{\rho}{2}, \frac{\rho}{2}\right], \quad \mathcal{C}_\sigma^{(1)} = \mathcal{C}_\sigma^{(0)} - \{(\sigma, 0)\}, \quad \mathcal{C}_\sigma^{(2)} = \mathcal{C}_\sigma^{(0)} + \{(\sigma, 0)\} \quad (0 < \sigma < \rho)$$

and let $\mathbb{P}$ be the mixture distribution over $\mathcal{X} = \mathcal{C}_\sigma^{(0)} \cup \mathcal{C}_\sigma^{(1)} \cup \mathcal{C}_\sigma^{(2)}$ given by

$$\mathbb{P} = \frac{1-\epsilon}{2}\Psi_1 + \frac{1-\epsilon}{2}\Psi_2 + \frac{\epsilon}{2}\Psi_0,$$

where Ψ_m is the uniform distribution over $\mathcal{C}_m$ for $m = 0, 1, 2$. The density function f of $\mathbb{P}$ is simply

$$f(x) = \frac{1}{\rho\sigma}\left(\frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(1)}) + \frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(2)}) + \frac{\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_\sigma^{(0)})\right) \tag{1}$$

so that for any $\epsilon < \lambda < (1-\epsilon)/2$, $\mathbb{C}_f(\lambda) = \left\{\mathcal{C}_\sigma^{(1)}, \mathcal{C}_\sigma^{(2)}\right\}$.

1.1 Lower bound for PPR.

As the following theorem demonstrates, even when Algorithm 1 is reasonably initialized, if the density cluster $\mathcal{C}_\sigma^{(1)}$ is sufficiently geometrically ill-conditioned the cluster estimator $\widehat{C}$ will fail to recover $\mathcal{C}_\sigma^{(1)}$. Let

$$\mathcal{L} = \{(x_1, x_2) \in \mathcal{X} : x_2 < 0\}. \tag{2}$$

Theorem 1. *Suppose* $r < \frac{1}{40}\rho \wedge \frac{1}{4}\sigma$, $\alpha = 65\Phi_{\mathbb{P}}(\mathcal{L})$, *and* $(L, U) = (0, 1)$ *are inputs to Algorithm 1. Then, for any*

$$n \geq \max\left\{\frac{64}{\epsilon^2\rho\sigma\pi r^2}, \frac{8}{\epsilon}\right\} \tag{3}$$

the following statement holds: there exists a set $C^g \subset X$ *with* $\mathrm{vol}(C^g \cap \mathcal{C}_\sigma^{(1)}[X]) \geq \frac{1}{4}\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X])$ *such that for any seed node* $v \in C^g$, *the estimator* $\widehat{C}$ *computed by Algorithm 1 has symmetric set difference with* $\mathcal{C}_\sigma^{(1)}[X]$ *of volume at least*

$$\frac{\sigma\rho}{r^2n^2}\frac{\mathrm{vol}(\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X])}{n} \geq \frac{1}{4} - c\frac{\sqrt{\frac{\sigma}{\rho}}}{\epsilon^2}\sqrt{\log\left(\frac{\rho\sigma}{\epsilon^2 r^2}\right)\frac{\sigma}{r}} \tag{4}$$

with probability at least $1 - c_1 n \exp\{-c_2 n\}$*, where* c *is a universal constant and* c_1, c_2 *are constants which do not depend on* n*.*

Roughly speaking, our lower bound implies that if $\sigma\sqrt{\frac{1}{r\rho}} \gg \epsilon^2$, then PPR will fail to recover the empirical density cluster $\mathcal{C}_1[X]$, for many seed nodes $v \in \mathcal{C}_1[X]$.

Theorem 1 is stated with respect to a particular hard case, where the density clusters are rectangular subsets of $\mathbb{R}^2$. Although we picked this setting to make the Theorem statement simple, our results are easily generalized to $\mathbb{R}^d$ and to non-rectangular clusters. Additionally, although we state our lower bound with respect to PPR run on neighborhood graph, the conclusion is likely to hold for a much broader class of spectral clustering algorithms. In the proof of Theorem 1, we rely heavily on the fact that when ϵ^2 is sufficiently greater than $\frac{\sigma}{\rho}$, the normalized cut of $\mathcal{C}_1$ will be much larger than that of $\mathcal{L}$. In this case, not merely PPR but any algorithm which approximates the minimum normalized cut is unlikely to recover $\mathcal{C}_1$. Local spectral clustering algorithms based on truncated random walks Spielman and Teng, global spectral clustering algorithms Shi and Malik, and p-Laplacian based spectral embeddings Buhler and Hein all have provable upper bounds on the normalized cut of cluster they output, and thus it stands to reason that all the aforementioned approaches should struggle to estimate $\mathcal{C}_1[X]$.

1.2 Comparison with upper bound.

To better digest the implications of Theorem 1, we apply our upper bound to the density function f given in (1).

Technically, this density does not fit nicely into the setup of Theorem 1, because the sets $\mathcal{C}_\sigma^{(m)}$ $(m = 1, 2)$ play the role of both the σ-expanded sets and the density clusters themselves. However, this truly is only a technical objection. One could modify the distribution $\mathbb{P}$ by adding a small amount of mass to the σ-interior $\mathcal{C}^{(m)} = \left\{x \in \mathcal{C}_\sigma^{(m)} : \text{dist}(x, \partial\mathcal{C}_\sigma) \geq \sigma\right\}$. Appropriately choosing a threshold $\lambda > \frac{(1-\epsilon)}{2\rho\sigma}$, the resulting λ-density clusters would be $\mathbb{C}_f(\lambda) = \left\{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}\right\}$, and their σ-expansions would resemble the rectangles $\mathcal{C}_\sigma^{(1)}, \mathcal{C}_\sigma^{(2)}$ except with slightly rounded corners. Alternatively, one could directly modify Lemma 6 – an auxiliary Lemma, used in the proof of Theorem 3, which establishes an upper bound on the Lebesgue measure of $\mathcal{C}_\sigma + B(0, r)$ – to hold under the assumption $\mathcal{C}_\sigma$ is a rectangle of width σ rather than a σ-expansion. The proof of this Lemma is the only time we use the fact that $\mathcal{C}_\sigma$ is a σ-expansion of a density cluster. (Ryan and Siva.)

Regardless, we will proceed to showing the behavior of our upper bound in this example. Observe that $\mathcal{C}_\sigma^{(1)}$ satisfies each of the assumptions (A1) - (A5). In particular

(A1) The density $f(x) = \frac{1-\epsilon}{2\rho\sigma}$ for all $x \in \mathcal{C}_\sigma^{(1)}$.

(A2) The clusters $\mathcal{C}_\sigma^{(1)}$ and $\mathcal{C}_\sigma^{(2)}$ are separated, $\text{dist}(\mathcal{C}_\sigma^{(1)}, \mathcal{C}_\sigma^{(2)}) \geq \sigma$.

(A3) The density $f(x) = \frac{\epsilon}{\rho\sigma}$ for all x such that $0 < \text{dist}(x, \mathcal{C}_\sigma^{(1)}) \leq \sigma$. Therefore for all such x,

$$\inf_{x' \in \mathcal{C}_\sigma^{(1)}} f(x') - f(x) > \left\{\frac{1-\epsilon}{2} - \epsilon\right\} \frac{1}{\rho\sigma}.$$

(A4) The set $\mathcal{C}_\sigma^{(1)}$ is itself convex, and has diameter ρ.

(A5) By symmetry, $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma^{(1)}) = \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma^{(2)})$ and therefore $\mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma^{(1)}) \leq \frac{1}{2}\mathrm{vol}_{\mathbb{P},r}(\mathbb{R}^d)$.

As long as the user-specified parameters are well-initialized, we may therefore apply Theorem 1. This implies that there exists a set $\mathcal{C}_\sigma^{(1)}[X] \subset \mathcal{C}_\sigma^{(1)}$ with $\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X]^g) \geq \frac{1}{2}\mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X])$ such that for any seed node $v \in \mathcal{C}_\sigma^{(1)}[X]$, the estimated cluster $\widehat{C}$ output by Algorithm 1 satisfies the following upper bound:

$$\mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X]) \leq c \cdot \kappa(\mathcal{C}^{(1)}) \cdot \mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X])$$

where the condition number is given by

$$\kappa(\mathcal{C}^{(1)}) = c\frac{\epsilon}{\sigma}\left(\frac{\rho^2}{r}\log^2\left(\frac{\Lambda_\sigma}{\lambda_\sigma^2 r}\right) + c\right).$$

Broadly speaking, our upper bound implies that when $\epsilon \ll \frac{\sigma r}{\rho^2}$, the volume of the symmetric set difference $\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X]$ will be small. To facilitate comparisons with our lower bound, assume $\frac{1}{4}\sigma \leq \frac{1}{40}\rho$ and set $r = \frac{1}{4}\sigma$. In this case, when $\epsilon \ll \frac{\sigma^2}{\rho^2}$ and the user-specified parameters satisfy (14), then with high probability the resulting PPR cluster estimator $\widehat{C}$ will have small symmetric set difference with $\mathcal{C}_\sigma^{(1)}[X]$. On the other hand, if $\epsilon \gg \left(\frac{\sigma}{\rho}\right)^{1/4}$ and the user-specified parameters satisfy the setup of Theorem 1, then with high probability $\widehat{C}$ will have large symmetric set difference with $\mathcal{C}_\sigma^{(1)}[X]$. Jointly, these upper and lower bounds give a relatively precise characterization of what it means for a density cluster to be well- or poorly-geometrically conditioned for recovery using PPR.

It is worth pointing out that both our lower and upper bounds are stated with respect to specific choices of the input parameters to Algorithm **??**, and that these choices are in fact different in the case of the teleportation parameter α and sweep cut range (L, U). At a high level, we do not believe this substantially weakens the takeaway messages of our work: if a density cluster $\mathcal{C}$ is geometrically well-conditioned, then a reasonable initialization of PPR will recover $\mathcal{C}_\sigma[X]$ with low error, whereas if it is geometrically poorly-conditioned, then a different but also reasonable initialization of PPR will fail to recover $\mathcal{C}_\sigma[X]$. Of course, it would be more satisfying if both our upper and lower bounds could be stated with respect to the same choices of these input parameters. Unfortunately, it is not at all obvious how to prove such a claim, as to the best of our knowledge all existing work on PPR assumes similar initialization conditions to the ones we assume; effectively, that the algorithm is well-initialized with respect to the set $S \subset G$ one is interested in recovering. We suspect that our lower bound continues to hold even when choosing input parameters as dictated by our upper bound; however, proving this is currently beyond our technical grasp, and is a matter for future work. Even so, we reiterate our belief that in tandem our upper and lower bounds represent a substantial step forward in characterizing the behavior of local clustering algorithms in a statistical context (Ryan and Siva).

1.3 Sample complexity.

- We stress that our lower bound characterizes the difficulty PPR has in identifying geometrically poorly-conditioned density clusters, but not the overall hardness of the density clustering problem in a statistical sense.

- It is a simple matter to prove that consistent estimation of the set $\mathcal{C}_\sigma$ is possible. (Ryan and Siva).

- To the best of our knowledge minimax rates for density clustering are known only when the density function is Lipschitz smooth, whereas in our setup the density f may not even be continuous. However, as a practical matter the extra cases included our setup are those where the clusters are very sharply defined, and therefore should be even easier to recover. See Rinaldo and Wasserman for a more thorough discussion on this matter.

- That PPR has difficulty recovering density clusters which can be easily estimated by standard plug-in approaches is not surprising, nor is it a knock on PPR. Rather, it simply reflects that while classical density clustering approaches are specifically designed to identify high-density regions regardless of their geometry, PPR considers geometry as well as density when deciding upon the optimal cluster.

2 Proof of Lower Bound.

To prove Theorem 1, we will proceeding according to the following steps:

1. We study the spectral partitioning properties of PPR on an arbitrary graph G, and show that when suitably initialized inside a subset $S \subset V$, the normalized cut of the PPR sweep cut is upper bounded by (a function of) $\Phi(S)$.

2. We specialize to the graph $G = G_{n,r}$ and the subset $\mathcal{L}[X] \subset X$, and show that the normalized cut $\Phi_{n,r}(\mathcal{L}[X])$ is small (with high probability) when the diameter ρ is large.

3. We reason that for the input parameters given in Theorem 1, the output of Algorithm 1 $\widehat{C}$ must therefore also have small normalized cut.

4. On other hand, we show that when the noise parameter ϵ is not too small, the empirical density cluster $\mathcal{C}_\sigma^{(1)}[X]$ will have large normalized cut $\Phi_{n,r}(\mathcal{C}_\sigma^{(1)}[X])$. In fact, we generalize this to hold for any set $A \subset X$ for which the symmetric set distance metric $\Delta(A, \mathcal{C}_1[X])$ is small.

5. We conclude that the symmetric set distance metric $\Delta(\widehat{C}, \mathcal{C}_\sigma^{(1)}[X])$ must not be small.

We devote the subsequent sections to proving each of the aforementioned steps.

2.1 Spectral partitioning properties of PPR.

Let $G = (V, E)$ be an undirected, unweighted graph with $m = |E|$ total edges, defined on vertices $V = \{v_1, \ldots, v_n\}$. Let C be a subset of the vertices V, Recall that for a given $\beta \in (0, 1)$ the sweep cut

$$S_{\beta,v} = \left\{ u \in V : \frac{p_v(u)}{\deg(u; G)} > \beta \right\}$$

The following theorem relates the normalized cut of the sweep sets $\Phi(S_\beta; G)$ to the normalized cut of C; it is stated with respect to the graph functionals

$$d_{\max} := \max_{u \in V} \deg(u; G), \quad \text{and} \quad d_{\min} := \min_{u \in V} \deg(u; G).$$

Theorem 2. *Let $C \subseteq V$ satisfy the following conditions:*

- $\mathrm{vol}(C; G) \leq \frac{2}{3}\mathrm{vol}(G)$,
- $|C| \geq \frac{d_{\max}}{d_{\min}}$, *and*
- $\frac{20\Phi(C;G)}{1+10\Phi(C;G)} + \frac{d_{\max}}{2d_{\min}^2} \leq \frac{1}{10}$.

Suppose $60\Phi(C; G) \leq \alpha \leq 70\Phi(C; G)$, and let $(L, U) = (0, 1)$. Then, there exists a subset $C^g \subset C$ with $\mathrm{vol}(C^g; G) \geq \frac{5}{6}\mathrm{vol}(C; G)$ such that for any $v \in C^g$ the following statement holds: For the PPR vector $p_v := p(v, \alpha; G)$, the minimum conductance sweep cut set satisfies

$$\min_{\beta \in (0,1)} \Phi(S_{\beta,v}; G) \leq \sqrt{11200 \left\{ \log\left(\frac{m}{d_{\min}^2}\right) + \log 20 \right\} \Phi(C; G)}$$

Although this theorem appears quite similar to standard results in the PPR literature – for instance, Theorem 6 of Anderson,Chung,Lang – crucially the above bound depends on $\log\left(\frac{m}{d_{\min}^2}\right)$ rather than $\log m$. In the case where $d_{\min} \asymp n$, this amounts to replacing a factor of $O(\log m)$ by a factor of $O(1)$, and therefore allows us to obtain meaningful results in the limit as $m \to \infty$.

Notwithstanding these improvements, the proof of Theorem 2 follows the same general outline as the proof of Theorem 6 of Anderson,Chung,Lang. We now walk through this outline step by step, modifying the results of Anderson,Chung,Lang as needed. As with their work, we begin by proving a mixing time bound on the PPR vector p_v.

2.1.1 Mixing time of PPR.

To quantify the mixing of a PPR vector p_v, we introduce the function $p[\cdot] : [0, 2m] \to [0, 1]$. For $j = 1, \ldots, n$, let β_j be the smallest value of $\beta \in (0, 1)$ such that S_{β_j} contains at least j vertices. (For notational ease, we will write $S_i := S_{\beta_i}$, so that $S_1, S_2, \ldots, S_n$ comprise the n unique sweep cuts of p_v.) For each $j = 1, \ldots, n$, we let $p[\mathrm{vol}(S_j)] = \sum_{u \in S} p_v(u)$. Additionally, we let $p[0] = 0$ and $p[2m] = 1$. Finally, we extend $p[\cdot]$ by piecewise interpolation to be defined everywhere on its domain. The mixedness of the PPR vector is then measured by the function $h : [0, 2m] \to [0, 1]$, defined as

$$h(k) = p[k] - \frac{k}{2m}.$$

Next, for a given $0 \le K_0 \le m$, let

$$L_{K_0}(k) = \frac{2m - K_0 - k}{2m - 2K_0} h(K_0) + \frac{k - K_0}{2m - 2K_0} h(2m - K_0)$$

be the linear interpolator of $h(K_0)$ and $h(2m - K_0)$, and additionally let

$$C(K_0) = \max\left\{ \frac{h(k) - L_{K_0}(k)}{\sqrt{\overline{k}}} : K_0 < k < 2m - K_0 \right\}.$$

where we use the notation $\overline{k} := \min\{k, 2m - k\}$.

Theorem 3 implies that if the PPR random walk is not well mixed, then some sweep cut of p_v must have small normalized cut.

Theorem 3. *Let $p_v = p(v, \alpha; G)$ be a PPR vector, and let ϕ be any constant in $[0, 1]$. Then, either the following bound holds for any integer t, any $0 < K_0 < m$, and any $k \in [K_0, 2m - K_0]$:*

$$h(k) \le \alpha t + L_{K_0}(k) + C(K_0)\sqrt{\overline{k}}\left(1 - \frac{\phi^2}{8}\right)^t \tag{5}$$

or else there exists some sweep cut S_j of p_v such that $\Phi(S_j; G) < \phi$.

Proof (of Theorem 3). The proof of Theorem 3 is essentially a combination of the proofs of Theorem 3 in Anderson,Chung,Lang and Theorem 1.2 in Lovasz and Simonovits. We will show that if $\Phi(S_j) > \phi$ for each $j = 1, \ldots, n$, then (5) holds for all t and any $k \in (K_0, 2m - K_0)$.

We proceed by induction on t. Our base case will be $t = 0$. Observe that $C(K_0) \cdot \sqrt{\overline{k}} \ge h(k) - L_{K_0}(k)$ for all $k \in [K_0, 2m - K_0]$, which implies

$$L_{K_0}(k) + C(K_0) \cdot \sqrt{\overline{k}} \ge h(k).$$

Now, we proceed with the inductive step. By the definition of L_{K_0}, the inequality (5) holds when $k = K_0$ or $k = 2m - K_0$. We will additionally show that (5) holds for every $k_j = \mathrm{vol}(S_j), j = 1, 2, \ldots, n$ such that $k_j \in [K_0, 2m - K_0]$. Once this is shown, the concavity of the expression on the right hands side of (5) implies that the inequality holds for all $k \in [K_0, 2m - K_0]$.

By Lemma 5 of Anderson,Chung,Lang, we have that

$$\begin{aligned} p[k_j] &\leq \alpha + \frac{1}{2}\left(p[k_j - |\partial(S_j)|] + p[k_j + |\partial S_j|]\right) \\ &\leq \alpha + \frac{1}{2}\left(p[k_j - \Phi(S_j)\overline{k}_j] + p[k_j + \Phi(S_j)\overline{k}_j]\right) \\ &\leq \alpha + \frac{1}{2}\left(p[k_j - \phi\overline{k}_j] + p[k_j + \phi\overline{k}_j]\right) \end{aligned}$$

and subtracting $k_j/2m$ from both sides, we get

$$h(k_j) \leq \alpha + \frac{1}{2}\big(h(k_j - \phi\overline{k}_j) + h(k_j + \phi\overline{k}_j)\big) \tag{6}$$

From this point, we divide our analysis into cases.

Case 1. Assume $k_j - 2\phi\overline{k}_j$ and $k_j + 2\phi\overline{k}_j$ are both in $[K_0, 2m - K_0]$. We are therefore in a position to apply our inductive hypothesis to (6), yielding

$$\begin{aligned} h(k_j) &\leq \alpha + \alpha(t-1)\frac{1}{2}\left(L_{K_0}(k_j - \phi\overline{k}_j) + L_{K_0}(k_j + \phi\overline{k}_j) + C(K_0)\big(\sqrt{\overline{k_j - \phi\overline{k}_j}} + \sqrt{\overline{k_j + \phi\overline{k}_j}}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right) \\ &\leq \alpha t + L_{K_0}(k) + \frac{1}{2}\left(C(K_0)\big(\sqrt{\overline{k_j - \phi\overline{k}_j}} + \sqrt{\overline{k_j + \phi\overline{k}_j}}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right) \\ &\leq \alpha t + L_{K_0}(k) + \frac{1}{2}\left(C(K_0)\big(\sqrt{\overline{k}_j - \phi\overline{k}_j} + \sqrt{\overline{k}_j + \phi\overline{k}_j}\big)\left(1 - \frac{\phi^2}{8}\right)^{t-1}\right). \end{aligned}$$

A Taylor expansion of $\sqrt{1+\phi}$ around $\phi = 0$ yields the following bound:

$$\sqrt{1+\phi} + \sqrt{1-\phi} \leq 2 - \frac{\phi^2}{4},$$

and therefore

$$h(k_j) \leq \alpha t + L_{K_0}(k) + \frac{C(K_0)}{2} \cdot \sqrt{\overline{k}_j} \cdot \left(2 - \frac{\phi^2}{4}\right)\left(1 - \frac{\phi^2}{8}\right)^{t-1} = \alpha t + L_{K_0}(k) + C(K_0)\sqrt{\overline{k}_j}\left(1 - \frac{\phi^2}{8}\right)^{t}.$$

Case 2.

Now, assume one of $k_j - 2\phi\overline{k}_j$ or $k_j + 2\phi\overline{k}_j$ is not in $[K_0, 2m - K_0]$. Without loss of generality assume $k_j < m$, so that (i) we have $k_j - 2\phi\overline{k}_j < K_0$ and (ii) $k_j + (k_j - K_0) \leq 2m - K_0$. By the concavity of h, and applying the inductive hypothesis to $h(2k_j - K_0)$, we have

$$\begin{aligned} h(k_j) &\leq \alpha + \frac{1}{2}\Big(h(K_0) + h\big(k_j + (k_j - K_0)\big)\Big) \\ &\leq \alpha + \frac{\alpha(t-1)}{2} + \frac{1}{2}\Big(L_{K_0}(K_0) + L_{K_0}(2k_j - K_0) + C(K_0)\sqrt{\overline{2k_j - K_0}}\left(1 - \frac{\phi^2}{8}\right)^{t-1}\Big) \\ &\leq \alpha t + L_{K_0}(k_j) + C(K_0)\frac{\sqrt{2\overline{k}_j}}{2}\left(1 - \frac{\phi^2}{8}\right)^{t-1} \\ &\leq \alpha t + L_{K_0}(k_j) + C(K_0)\sqrt{\overline{k}_j} \cdot \left(1 - \frac{\phi^2}{8}\right)^{t} \end{aligned}$$

□

As a sanity check, we confirm that Theorem 3 is no weaker than Theorem 3 of Anderson, Chung,Lang. It is not hard to show that $h(k) \leq \min\{1, \sqrt{k}\}$, and therefore that $C(K_0) \leq 1$ for any K_0. Setting $K_0 = 0$ in Theorem 3, we therefore recover Theorem 3 of Anderson, Chung,Lang.

We now proceed to identify when Theorem 3 may offer some improvement on Theorem 3 of Anderson, Chung,Lang, by showing when we can upper bound $C(K_0) << 1$. The critical point is that since $h(k)$ is concave and $L_{K_0}(K_0) = h(K_0)$ the upper bound

$$\frac{h(k) - L_{K_0}(k)}{\sqrt{k}} \leq h'(K_0)\sqrt{k}$$

holds whenever $k < m$. For similar reasons, when $k > m$,

$$\frac{h(k) - L_{K_0}(k)}{\sqrt{k}} \leq -h'(2m - K_0)\sqrt{2m - k}.$$

(Since h is not differentiable at points $k = \mathrm{vol}(S_j)$, here we use h' to denote the left derivative of h whenever $k < m$, and the right derivative of h whenever $k \geq m$)

The following Lemma gives good estimates for $h'(K_0)$ and $h'(2m - K_0)$, and a resulting upper bound on $C(K_0)$.

Lemma 1. *There exists $K_0 \in \{0, \deg(v; G)\}$ such that*

$$h'(K_0) \leq \frac{1}{2d_{\min}^2}. \tag{7}$$

Additionally, for all $K_0 \in [0, 2m]$,

$$h'(2m - K_0) \geq -\frac{d_{\max}}{d_{\min}\mathrm{vol}(G)}. \tag{8}$$

As a result,

$$C(K_0) \leq \frac{\sqrt{m}}{d_{\min}^2}.$$

Proof (of Lemma 1). The result of the Lemma is obvious once we show (7) and (8). To show either inequality, it will be useful to work with an alternative representation of h. In particular, whenever $\mathrm{vol}(S_j) \leq k < \mathrm{vol}(S_{j+1})$ (where we let $S_0 = \emptyset$), the function $h(k)$ may be written as

$$h(k) = \sum_{i=0}^{j} \left(p_v(u_{(i)}) - \pi(u_{(i)}; G)\right) + \frac{\left(k - \mathrm{vol}(S_j; G)\right)}{\deg(u_{(j+1)}; G)} \left(p_v(u_{(j+1)}) - \pi(u_{(j+1)}; G)\right) \tag{9}$$

where the vertices are ordered $\frac{p_v(u_{(1)})}{\deg(u_{(1)};G)} \geq \frac{p_v(u_{(2)})}{\deg(u_{(2)};G)} \geq \cdots \geq \frac{p_v(u_{(n)})}{\deg(u_{(n)};G)}$, and as usual $\pi(u; G) = \frac{\deg(u;G)}{\mathrm{vol}(G)}$.

From this representation, it is not hard to verify that the left derivative $h'(k)$ can be upper bounded

$$h'(k) \leq \frac{p(v_{(j+1)})}{\deg(v_{(j+1)}; G)} \tag{10}$$

We now upper bound $p(u)$ uniformly over all u except the seed node v. For any $u \in V$ besides the seed node v, we can show by induction that

$$e_v W^t(u) \leq \frac{1}{2d_{\min}}$$

for any $t \geq 0$, and therefore

$$p(\alpha, \chi_v)(u) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \chi_v W^t(u) \leq \frac{1}{2d_{\min}}. \tag{11}$$

As a result, by (10), for either $K_0 = \deg(v; G)$ (in the case where $v_{(1)} = v$) or otherwise for $K_0 = 0$, the inequality $h'(K_0) \leq \frac{1}{2d_{\min}^2}$ holds, proving (7). The inequality (8) follows immediately from the representation (9), since

$$h'(k) \geq -\frac{\pi(v_{(j+1)})}{d(v_{(j+1)})} \geq -\frac{\pi_{\max}}{d_{\min}},$$

and the proof of the Lemma is therefore complete. □

To apply Theorem 3, we must also upper bound the linear interpolator $L_{K_0}(k)$. Of course, trivially $L_{K_0}(k) \leq \max\{h(K_0), h(2m - K_0)\}$ for all k. As it happens, this observation will lead to a sufficient upper bound on L_{K_0}.

Lemma 2. *Assume $s = \chi_v$ for some $v \in V$. Let $K_0 = \mathrm{vol}(S_j)$ for some $j = 0, \ldots, n$. Then,*

$$h(2m - K_0) \leq \frac{K_0}{2m} \text{ and } h(K_0) \leq \frac{K_0}{2d_{\min}^2} + \frac{2\alpha}{1+\alpha}.$$

and as a result for any $k \in \mathbb{R}$,

$$L_{K_0}(k) \leq \frac{2\alpha}{1+\alpha} + \frac{K_0}{2d_{\min}^2}.$$

Proof (of Lemma 2). We make use of the representation (9) to prove the desired upper bounds on $h(2m - K_0)$ and $h(K_0)$. We first upper bound $h(2m - K_0)$,

$$\begin{aligned} h(2m - K_0) &= \sum_{i=1}^{j} p(v_{(i)}) - \pi(v_{(i)}) \\ &\leq 1 - \sum_{i=1}^{j} \pi(v_{(i)}) \\ &= 1 - \sum_{i=1}^{j} \frac{d(v_i)}{2m} = \frac{K_0}{2m}. \end{aligned}$$

We will upper bound $h(K_0)$ by $p[\mathrm{vol}(S_j)] \leq p_v(v) + \sum_{u \in S_j \setminus \{v\}} p_v(u)$. In the proof of Lemma 1 we have already given an upper bound on $p_v(u)$ when $u \neq v$. Now, we additionally observe that for all t,

$$e_v W^t(v) \leq \frac{1}{2d_{\min}} + \left(\frac{1}{2}\right)^t$$

and therefore $p_v(v) \leq \frac{1}{2d_{\min}} + \frac{2\alpha}{1+\alpha}$. As a result,

$$h(K_0) \leq \frac{2\alpha}{1+\alpha} + \frac{|S_j|}{2d_{\min}} \leq \frac{2\alpha}{1+\alpha} + \frac{K_0}{2d_{\min}^2}, \tag{12}$$

where the latter inequality follows since $K_0 = \mathrm{vol}(S_j) \geq |S_j| \cdot d_{\min}$. □

Combining Theorem 3, Lemma 1 and Lemma 2, we have the following result.

Corollary 1. *Let $p_v = p(v, \alpha; G)$ be a PPR vector with seed node $v \in V$, and let ϕ be any constant in $[0, 1]$. Then, either the following bound holds for any integer t and any $k \in [d_{\max}, 2m - d_{\min}]$:*

$$h(k) \leq \alpha t + \frac{2\alpha}{1+\alpha} + \frac{d(v)}{2d_{\min}^2} + \frac{\sqrt{m}}{d_{\min}^2} \cdot \sqrt{k} \left(1 - \frac{\phi^2}{8}\right)^t$$

or there exists some sweep cut S_j of p_v such that $\Phi(S_j; G) < \phi$.

We arrive now at the main result of this section. It is similar in form to Theorem 2 of Anderson,Chung,Lang but reflects the improvements due to using Corollary 1. To simplify notation, we will write the total mass placed by p_v on a subset $S \subset V$ as $p_v(S) := \sum_{u \in S} p_v(u)$.

Theorem 4. *Let $p_v = p(v, \alpha; G)$ be a PPR vector with seed node $v \in V$. Suppose there exists some $\delta > \frac{2\alpha}{1+\alpha} + \frac{d_{\max}}{2d_{\min}^2}$, such that*

$$p_v(S) - \frac{\mathrm{vol}(S; G)}{\mathrm{vol}(G)} > \delta \tag{13}$$

for a set S with cardinality $|S| \geq \frac{d_{\max}}{d_{\min}}$. Then there exists a sweep cut S_j of p, such that

$$\Phi(S_j) < \sqrt{\frac{16\alpha \left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log\left(\frac{2}{\delta'}\right)\right\}}{\delta'}}$$

where $\delta' = \delta - \frac{2\alpha}{1+\alpha} + \frac{d(v)}{2d_{\min}^2}$.

Proof. Suppose the assumption of the theorem is satisified, that is there exists a set $S \subset V$ with cardinality $|S| \geq \frac{d_{\max}}{d_{\min}}$ which satisfies (13). Then for $j = |S|$ the sweep cut S_j has volume at least $d_{\max}$, and by hypothesis $h(\mathrm{vol}(S_j)) > \delta$.

Now, letting

$$t = \frac{8}{\phi^2}\left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log\left(\frac{2}{\delta'}\right)\right\}, \quad \phi^2 = \frac{16\alpha \left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log(\frac{2}{\delta'})\right\}}{\delta'}$$

we have that

$$\alpha t + \frac{2\alpha}{1+\alpha} + \frac{d(v)}{2d_{\min}^2} + \frac{\sqrt{m}}{d_{\min}^2} \cdot \sqrt{k}\left(1 - \frac{\phi^2}{8}\right)^t \leq \frac{\delta'}{2} + \frac{2\alpha}{1+\alpha} + \frac{d(v)}{2d_{\min}^2} + \frac{\delta'}{2} < \delta,$$

and the Theorem follows by Corollary 1. □

2.1.2 Improved Local Partitioning with PPR.

As in Anderson,Chung,Lang, the mixing time results of the previous section lead to an upper bound on the normalized cut $\Phi(\widehat{C}; G)$. First, we restate a theorem of Anderson,Chung,Lang which lower bounds the probability mass $p(v, \alpha; G)(C)$ as a function of the normalized cut $\Phi(C)$.

Theorem 5. *For any set C and any constant α, there exists a subset $C^g \subset C$ with $\mathrm{vol}(C^g; G) \geq \frac{5}{6}\mathrm{vol}(C; G)$, such that for any vertex $v \in C^g$, the PPR vector $p(v, \alpha; G)$ satisfies*

$$p(v, \alpha; G) \geq 1 - 6\frac{\Phi(C; G)}{\alpha}.$$

We are now in a position to prove Theorem 2 by combining Corollary 1 and Theorem 5.

Proof (of Theorem 2). Since $\alpha \geq 60\Phi(C)$ and $v \in C^g$, by Theorem 5,

$$p_v(C) \geq \frac{9}{10}.$$

This inequality along with the assumption $\mathrm{vol}(C) \leq \frac{2}{3}\mathrm{vol}(G)$ implies that $p_v(C) - \frac{\mathrm{vol}(C)}{\mathrm{vol}(G)} \geq \frac{1}{5}$. Since we assume $|C| \geq \frac{d_{\max}}{d_{\min}}$, the hypothesis of Theorem 4 is satisfied with $\delta = 1/5$. Therefore, the minimum conductance sweep cut satisfies

$$\min_{j=1,\ldots,n} \Phi(S_j; G) \leq \sqrt{\frac{1120 \cdot \Phi(C; G)\left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log\left(\frac{2}{\delta'}\right)\right\}}{\delta'}}$$

Finally, we assume $\frac{20\Phi(C)}{1+10\Phi(C)} + \frac{d_{\max}}{2d_{\min}^2} \leq \frac{1}{10}$ which implies that

$$\delta' = \delta - \frac{20\alpha}{1+10\alpha} + \frac{d_{\max}}{2d_{\min}^2} \geq \frac{1}{10}$$

completing the proof of the theorem. □

2.2 Normalized cut of $\mathcal{L}[X]$.

Recall that for any set $\mathcal{A} \subset \mathcal{X}$, the $\mathbb{P}$-weighted *cut* and *volume* functionals can be written as

$$\mathrm{cut}_{\mathbb{P},r}(\mathcal{A}) = \int_{\mathcal{A}}\int_{\mathcal{X}\setminus\mathcal{A}} \mathbf{1}(\|x-y\| \leq r)\, d\mathbb{P}(x)\, d\mathbb{P}(y), \mathrm{vol}_{\mathbb{P},r}(\mathcal{A}) := \int_{\mathcal{A}}\int_{\mathcal{X}} \mathbf{1}(\|x-y\| \leq r)\, d\mathbb{P}(x)\, d\mathbb{P}(y),$$

and the continuous *normalized cut* is

$$\Phi_{\mathbb{P},r}(\mathcal{A}) := \frac{\mathrm{cut}_{\mathbb{P},r}(\mathcal{A})}{\min\{\mathrm{vol}_{\mathbb{P},r}(\mathcal{A}), \mathrm{vol}_{\mathbb{P},r}(\mathcal{X}\setminus\mathcal{A})\}}.$$

We now upper bound the normalized cut $\Phi_{\mathbb{P},r}(\mathcal{L})$ as a function of the diameter ρ, and the neighborhood graph radius r. Our bounds will be simple and not tight, but will display the right dependence on these parameters, and so will be sufficient for our purposes.

To upper bound $\mathrm{cut}_{\mathbb{P},r}(\mathcal{L})$, note that for any $x = (x_1, x_2) \in \mathcal{L}$, if $x_2 \leq -r$ the ball $B(x, r)$ and the set $\mathcal{X} \setminus \mathcal{L}$ are disjoint. This implies

$$\begin{aligned}\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}) &\leq \mathbb{P}(\{x \in \mathcal{X} : -r < x_2 < 0\}) \cdot \max_{x\in\mathcal{X}} \mathbb{P}(B(x, r)) \\ &\leq \frac{r}{2\rho} \cdot \frac{\pi r^2}{2\sigma\rho}.\end{aligned}$$

By symmetry, $\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) = \mathrm{vol}_{\mathbb{P},r}(\mathcal{X} \setminus \mathcal{L})$, and therefore to upper bound $\Phi_{\mathbb{P},r}(\mathcal{L})$, it is sufficient to lower bound $\mathrm{vol}_{\mathbb{P},r}(\mathcal{L})$. We have

$$\begin{aligned}\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) &\geq \mathbb{P}(\{x \in \mathcal{C}_1 \cap \mathcal{L} : \mathrm{dist}(x, \partial\mathcal{C}_1) > r\}) \cdot \frac{\pi r^2}{2\sigma\rho} \\ &= \frac{(\sigma - 2r)(\rho - r)}{2\sigma\rho} \cdot \frac{\nu_d r^d}{2\sigma\rho} \\ &\geq \frac{3}{16} \cdot \frac{\pi r^2}{2\sigma\rho}\end{aligned}$$

where the last inequality follows since $r \leq \frac{1}{4}\sigma < \frac{1}{4}\rho$. Therefore, $\Phi_{\mathbb{P},r}(\mathcal{L}) \leq \frac{8r}{3\rho}$.

Then, Lemma 4 implies that the graph functionals $\mathrm{cut}_{n,r}(\mathcal{L}[X])$ and $\mathrm{vol}_{n,r}(\mathcal{L}[X])$–and in turn $\Phi_{n,r}(\mathcal{L}[X])$– concentrate around their expectations. Precisely, we have that

$$\begin{aligned}\Phi_{n,r}(\mathcal{L}[X]) &= \frac{\mathrm{cut}_{n,r}(\mathcal{L}[X])}{\min\{\mathrm{vol}_{n,r}(\mathcal{L}[X]), \mathrm{vol}_{n,r}((\mathcal{X} \setminus \mathcal{L})[X])\}} \\ &\leq \frac{3}{2}\Phi_{\mathbb{P},r}(\mathcal{A}) \leq \frac{4r}{\rho} \qquad (14)\end{aligned}$$

with probability at least $1 - 3\exp\{-\frac{1}{25}n(\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}))^2\}$.

2.3 Normalized cut of $\widehat{C}$.

We will use Lemma 4 to show that the set $\mathcal{L}[X] \subset X$ and the neighborhood graph $G_{n,r}$ satisfy the required conditions of Theorem 2. This Theorem will imply an upper bound on the normalized cut of $\widehat{C}$.

Note that $\mathcal{X}$ and f satisfy the regularity conditions (A) and (B), with parameters $\lambda_{\min} = \frac{\epsilon}{\rho\sigma}, \lambda_{\max} = \frac{1}{2\rho\sigma}$, and $a = 4, R = \frac{1}{4}\sigma$. We therefore have that for any $\delta \in (0,1)$ and any $r \in (0, \frac{1}{4}\sigma)$, each of the following inequalities are satisfied with probability at least $1 - 2n\exp\left\{-\frac{\pi\epsilon r^2\delta^2 n}{8\rho\sigma(1+\frac{\delta}{3})}\right\} - 6\exp\left\{-n\delta^2(\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}[X]))^2\right\}$:

- $\frac{(1-\delta)\epsilon\pi r^2}{4\rho\sigma}n \leq d_{\min} \leq d_{\max} \leq \frac{(1+\delta)\pi r^2}{2\rho\sigma}n$,
- $\frac{(1-\delta)}{2}n \leq |\mathcal{L}[X]| \leq \frac{(1+\delta)}{2}n$,
- $\mathrm{vol}_{n,r}(\mathcal{L}[X]) \leq (1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{L}) = \frac{(1+\delta)}{2}\mathrm{vol}_{\mathbb{P},r}(\mathcal{X}) \leq \frac{(1+\delta)}{2}\mathrm{vol}(G_{n,r})$, and
- $(1-\delta)\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}[X]) \leq \mathrm{cut}_{n,r}(\mathcal{L}[X]) \leq (1+\delta)\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}[X])$.

We now condition on these inequalities, and letting $\delta = \frac{2}{67}$ we verify that under the setup of Theorem 1, each of the conditions of Theorem 2 are met:

- $\mathrm{vol}(\mathcal{L}[X]) \leq \frac{(1+\delta)}{2(1-\delta)}\mathrm{vol}(G_{n,r}) \leq \frac{2}{3}\mathrm{vol}(G_{n,r})$ since $\delta < 1/7$,
- $|\mathcal{L}[X]| \geq \frac{n(1-\delta)}{2} \geq \frac{2(1+\delta)}{(1-\delta)\epsilon} \geq \frac{d_{\max}}{d_{\min}}$ and $\frac{d_{\max}}{2d_{\min}^2} \leq \frac{8(1+\delta)}{(1-\delta)^2\epsilon^2\rho\sigma\pi r^2} \cdot \frac{1}{n} \leq \frac{1}{10}$ by (3),
- $\Phi_{n,r}(\mathcal{L}[X]) \leq \frac{4r}{\rho} \leq \frac{1}{10}$, by assumption on r and ρ, and
- $60\Phi_{n,r}(\mathcal{L}[X]) \leq \frac{60(1+\delta)}{1-\delta}\Phi_{\mathbb{P},r}(\mathcal{L}) \leq \alpha \leq \frac{65(1+\delta)}{1-\delta}\Phi_{n,r}(\mathcal{L}[X]) \leq 70\Phi_{n,r}(\mathcal{L}[X])$ since $\delta < 2/67$.

We may therefore apply Theorem 2, which allow us to upper bound the minimum conductance sweep cut $\min_{\beta\in(0,1)}\Phi(S_{\beta,v};G)$ or equivalently the output of Algorithm 1.

To be precise, we have that there exists a set $\mathcal{L}[X]^g \subset \mathcal{L}[X]$ with $\mathrm{vol}_{n,r}(\mathcal{L}[X]^g) \geq \frac{5}{6}\mathrm{vol}_{n,r}(\mathcal{L}[X])$, such that the following statement holds for any $v \in \mathcal{L}[X]^g$: when Algorithm 1 is run with inputs $X, r < \frac{1}{4}\sigma, \alpha = 65\Phi_{\mathbb{P},r}(\mathcal{L}[X]), v \in \mathcal{L}[X]^g$ and $(L,U) = (0,1)$, the resulting PPR cluster estimate $\widehat{C}$ satisfies

$$\begin{aligned}\Phi_{n,r}(\widehat{C}) &\leq \sqrt{11200\left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log 20\right\}\Phi_{n,r}(\mathcal{L}[\mathcal{X}])} \\ &\leq \sqrt{89600\left\{\log\left(\frac{\rho\sigma}{\epsilon^2\pi r^2}\right) + \log 20\right\}\frac{r}{\rho}} \end{aligned} \tag{15}$$

with probability at least $1 - 2n\exp\left\{-\frac{\pi\epsilon r^2 n}{8978\rho\sigma}\right\} - 6\exp\left\{-\frac{1}{1123}(\mathrm{cut}_{\mathbb{P},r}(\mathcal{L}[X]))^2 n\right\}$ (where the latter inequality follows from (14) and Lemma 4.)

2.4 Lower bound on normalized cut.

The precise statement we will prove is contained in the following Lemma.

Lemma 3. *The normalized cut $\Phi_{n,r}(A)$ is upper bounded*

$$\Phi_{n,r}(A) \geq \frac{1}{12\pi}\left(1 - 4\frac{\sigma\rho}{r^2 n^2}\mathrm{vol}_{n,r}(A \triangle \mathcal{C}_\sigma^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma}$$

uniformly over all $A \subset X$ with probability at least $1 - c_1\exp\{-c_2 n\}$.

Proof. To lower bound the normalized cut $\Phi_{n,r}(A)$, we must lower bound $\mathrm{cut}_{n,r}(A)$ and upper bound $\mathrm{vol}_{n,r}(A)$. A naive upper bound on the volume is simply

$$\mathrm{vol}_{n,r}(A) \leq \mathrm{vol}_{n,r}(G_{n,r}) \leq (1+\delta)\mathrm{vol}_{\mathbb{P},r}(\mathcal{X})n^2 \leq (1+\delta)\frac{\pi r^2}{\rho\sigma}n^2 \tag{16}$$

where the last inequality holds with probability at least $1 - c_1\exp\{-c_2 n\}$, and it turns out this will suffice for our purposes. (Here and in the rest of this proof we take $\delta = 1/2$.)

We turn to lower bounding $\mathrm{cut}_{n,r}(A)$. We will approximate the cut of A by discretizing the space $\mathcal{X}$ into bins, relate the cut of A to the boundary of the binned set $\overline{A}$, and then lower bound the size of the boundary of $\overline{A}$.

Let (k_1, k_2) for $k_1 \in \left[\frac{6\sigma}{r}\right], k_2 \in \left[\frac{2\rho}{r}\right]$ be the upper right hand corner of the cube

$$Q_{(k_1,k_2)} = \left[-\frac{3\sigma}{2} + \frac{(k_1-1)}{2}r, -\frac{3\sigma}{2} + \frac{k_1}{2}r\right] \times \left[-\frac{\rho}{2} + \frac{(k_2-1)}{2}r, -\frac{\rho}{2} + \frac{k_2}{2}r\right]$$

and let $\overline{Q} = \left\{Q_{(k_1,k_2)} : k_1 \in \left[\frac{6\sigma}{r}\right], k_2 \in \left[\frac{2\rho}{r}\right]\right\}$ be the collection of such cubes. For a set $A \subset X$ we define the binned set $\overline{A} \subset \overline{Q}$ as follows

$$\overline{A} := \left\{Q \in \overline{Q} : \mathbb{P}_n(A \cap Q) \geq \frac{1}{2}\mathbb{P}_n(Q)\right\},$$

and we let

$$\partial\overline{A} := \left\{Q_{(k_1,k_2)} \in \overline{A} : \exists(\ell_1,\ell_2) \in \left[\frac{3\sigma}{r}\right] \times \left[\frac{\rho}{r}\right] \text{ such that } Q_{(\ell_1,\ell_2)} \notin \overline{A}, \|k-\ell\|_1 = 1\right\}.$$

be the boundary set of $\overline{A}$ in $\overline{Q}$. Intuitively, every point $x_i \in A$ in the boundary set of $\overline{A}$ will have many edges to $X \setminus A$. Formally, letting $Q_{\min} := \min_{Q \in \overline{Q}} \mathbb{P}_n(Q)$, we have

$$\mathrm{cut}_{n,r}(A) \geq \mathrm{cut}_{n,r}(A \cap \{x_i \in \overline{A}\}) \geq \frac{1}{4} \left|\partial \overline{A}\right| Q_{\min}^2, \tag{17}$$

where the last inequality follows since for every cube $Q_k \in \partial\overline{A}$, there exists a cube $Q_\ell \notin \overline{A}$ such that $\|i - j\|_1 \leq 1$, and since each cube has side length $r/2$, this implies that for every $x_i \in Q_k$ and $x_j \in Q_\ell$ the edge (x_i, x_j) belongs to $G_{n,r}$.

Now we move on lower bounding the size of the boundary $\left|\partial\overline{A}\right|$. To do so, we divide $\mathcal{X}$ into slices horizontally. Let $R_k = \left\{(x_1, x_2) \in \mathcal{X} : x_2 \in \left[-\frac{\rho}{2} + \frac{(k-1)}{2}r, -\frac{\rho}{2} + \frac{k}{2}r\right]\right\}$ be the kth horizontal slice, and $\overline{R}_k = \left\{Q_{(k_1,k)} \in \overline{Q} : k_1 \in [\frac{6\sigma}{r}]\right\}$ be the binned version of R_k. For each k, either

1. $\overline{R}_k \cap \overline{A} = \emptyset$, in which case

$$\mathrm{vol}_{n,r}\Big(\big(A \vartriangle C_1[X]\big) \cap R_k\Big) \geq \frac{1}{2}\mathrm{vol}_{n,r}(C_1[X] \cap R_k), \quad \text{or}$$

2. $\overline{R}_k \cap \overline{A} = \overline{R}_k$, in which case

$$\mathrm{vol}_{n,r}\Big(\big(A \vartriangle C_1[X]\big) \cap R_k\Big) \geq \frac{1}{2}\mathrm{vol}_{n,r}(C_2[X] \cap R_k), \quad \text{or}$$

3. $\overline{R}_k \cap \partial\overline{A} \neq \emptyset$.

Let $N(R)$ be the number of slices for which $\overline{R}_k \cap \partial\overline{A} \neq \emptyset$. By the cases elucidated above, letting

$$R_{\min} := \min_k \Big\{\mathrm{vol}_{n,r}(C_1[X] \cap R_k) \wedge \mathrm{vol}_{n,r}(C_2[X] \cap R_k)\Big\}$$

we obtain the following lower bound on the volume of the symmetric set difference,

$$\mathrm{vol}_{n,r}(A \vartriangle C_1[X]) \geq \frac{1}{2} R_{\min} \left[\frac{2\rho}{r} - N(R)\right]. \tag{18}$$

Finally note that $\left|\partial\overline{A}\right| \geq N(R)$. Therefore combining (17) and (18), we have that

$$\begin{aligned}\mathrm{cut}_{n,r}(A) &\geq \frac{1}{4} N(R) Q_{\min}^2 \\ &\geq \frac{1}{2}\left(\frac{\rho}{r} - \frac{\mathrm{vol}_{n,r}(A \vartriangle C_1[X])}{R_{\min}}\right) Q_{\min}^2 \end{aligned} \tag{19}$$

for all $A \subset X$.

It remains to lower bound the random quantities $R_{\min}$ and $Q_{\min}$. To do so, we first lower bound the expected probability of any cell Q,

$$\min_{Q \in \overline{Q}} \mathbb{P}(Q) \geq \frac{\epsilon r^2}{\rho\sigma}.$$

and the expected volume of $\mathcal{C}_\sigma^{(1)}[X] \cap R_k$ and $\mathcal{C}_\sigma^{(2)}[X] \cap R_k$,

$$\mathbb{E}(\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(2)}[X] \cap R_k)) = \mathbb{E}(\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X] \cap R_k)) = \mathrm{vol}_{\mathbb{P},r}(\mathcal{C}_\sigma^{(1)} \cap R_k) \geq \frac{1}{2}\frac{\pi r^3}{\sigma\rho^2}$$

Since $Q_{\min}$ and $R_{\min}$ are obtained by taking the minimum of functionals over a fixed number of sets in n, applying Bernstein's inequality and a union bound gives

$$Q_{\min} \geq (1-\delta)\frac{\epsilon r^2}{\rho\sigma} \quad \text{and} \quad R_{\min} \geq \frac{(1-\delta)}{2}\frac{\pi r^3}{\sigma\rho^2},$$

with probability at least $1 - c_1\exp\{-c_2 n\}$. Combining these lower bounds with (16) and (19), we obtain

$$\Phi_{n,r}(A) \geq \frac{(1-\delta)^2}{2(1+\delta)\pi}\left(1 - 2\frac{\sigma\rho}{(1-\delta)r^2n^2}\mathrm{vol}_{n,r}(A \triangle \mathcal{C}_\sigma^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma}$$

with probability at least $1 - c_1\exp\{-c_2 n\}$. □

Conclusion. Combining (15) and Lemma 3, we have that there exists a set $\mathcal{L}[X]^g \subset \mathcal{L}[X]$ with $\mathrm{vol}_{n,r}(\mathcal{L}[X]^g) \geq \frac{5}{6}\mathrm{vol}_{n,r}(\mathcal{L}[X])$ such that for any seed node $v \in \mathcal{L}[X]^g$, the following bounds hold:

$$\frac{1}{12\pi}\left(1 - 4\frac{\sigma\rho}{r^2n^2}\mathrm{vol}_{n,r}(\widehat{C} \triangle \mathcal{C}_\sigma^{(1)}[X])\right)\frac{\epsilon^2 r}{\sigma} \leq \Phi_{n,r}(\widehat{C}) \leq \sqrt{89600\left\{\log\left(\frac{\rho\sigma}{\epsilon^2\pi r^2}\right) + \log 20\right\}\frac{r}{\rho}},$$

with probability at least $1 - c_1\exp\{-c_1 n\}$. Finally, we show that the volume of $\mathcal{L}[X]^g$ is sufficiently large to ensure that it includes many points in $\mathcal{C}_\sigma^{(1)}[X]$:

$$\begin{aligned}
\mathrm{vol}_{n,r}(\mathcal{L}[X]^g \cap \mathcal{C}_\sigma^{(1)}[X]) &\geq \mathrm{vol}_{n,r}(\mathcal{L}[X]^g) - \mathrm{vol}_{n,r}((\mathcal{L}[X]^g \cap (\mathcal{C}_\sigma^{(0)} \cup \mathcal{C}_\sigma^{(2)})[X])[X]) \\
&\geq \frac{5}{6}\mathrm{vol}_{n,r}(\mathcal{L}[X]) - \mathrm{vol}_{n,r}((\mathcal{L} \cap (\mathcal{C}_\sigma^{(0)} \cup \mathcal{C}_\sigma^{(1)})[X])[X]) \\
&\geq \frac{5}{6}\mathrm{vol}_{n,r}((\mathcal{L} \cap \mathcal{C}_\sigma^{(1)})[X]) - \frac{1}{6}\mathrm{vol}_{n,r}(\mathcal{L}[X]) \\
&\geq \left(\frac{5}{6} - \frac{(1+\delta)}{2(1-\delta)}\right)\mathrm{vol}_{n,r}((\mathcal{L} \cap \mathcal{C}_\sigma^{(1)})[X]) \\
&\geq \frac{(1-\delta)}{2(1+\delta)}\left(\frac{5}{6} - \frac{(1+\delta)}{2(1-\delta)}\right)\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X])
\end{aligned}$$

with high probability, where the final two inequalities follow from Lemma 4. Setting $\delta = 1/13$, we have that $\mathrm{vol}_{n,r}(\mathcal{L}[X]^g \cap \mathcal{C}_\sigma^{(1)}[X]) \geq \frac{1}{4}\mathrm{vol}_{n,r}(\mathcal{C}_\sigma^{(1)}[X])$.

3 Concentration Inequalities.

Let f be a density function, defined on the domain $\mathcal{X}$. In our theory, we frequently appeal to concentration of degree and volume graph functionals around their means. In this section, we establish that this concentration holds under certain regularity conditions on f. In particular, we will assume

(A) There exist $\lambda_{\min}$ and $\lambda_{\max}$ such that for any $x \in \mathcal{X}$:

$$0 < \lambda_{\min} < f(x) < \lambda_{\max} < \infty.$$

(B) There exists some $a, R > 0$ such that for any $0 < r < R$ and any $x \in \mathcal{X}$,

$$\nu(B(x,r) \cap \mathcal{X}) \geq \frac{\nu_d r^d}{a},$$

where $\nu(\cdot)$ is Lebesgue measure on $\mathbb{R}^d$ and $B(x,r)$ is the ball centered at x of radius r.

We collect our bounds on graph functionals in the following Lemma.

Lemma 4. *Let $X = \{x_1, \dots, x_n\}$ be sampled independently from $\mathbb{P}$, let $G := G_{n,r}$ be a neighborhood graph over X, and let $\mathcal{S}$ be a subset of $\mathcal{X}$. Suppose the density f of $\mathbb{P}$ satisfies the regularity conditions (A) and (B). Then, for any $\delta \in (0,1)$ and any $r \in (0, R)$, there exists numbers $c_1 := c_1(f)$ and $c_2 := c_2(f)$ independent of the sample size n such that each of the following bounds hold.*

With probability at least $1 - 2n \exp\left\{ -\frac{\frac{1}{2}\delta^2 \frac{\lambda_{\min} \nu_d r^d}{a} n}{1 + \frac{\delta}{3}} \right\}$,

$$(1-\delta)\lambda_{\min} \frac{\nu_d r^d}{a} n \leq d_{\min} \leq d_{\max} \leq (1+\delta)\lambda_{\min} \frac{\nu_d r^d}{a} n.$$

With probability at least $1 - 2\exp\left\{-2n\delta^2 \mathbb{P}(S)^2\right\}$,

$$(1-\delta)\mathbb{P}(\mathcal{S})n \leq |\mathcal{S}[X]| \leq (1-\delta)\mathbb{P}(\mathcal{S})n.$$

With probability at least $1 - 2\exp\left\{-n\delta^2 \mathrm{vol}_{\mathbb{P},r}(\mathcal{S})^2\right\}$,

$$(1-\delta)n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}) \leq \mathrm{vol}_{n,r}(\mathcal{S}[X]) \leq (1+\delta)n(n-1)\mathrm{vol}_{\mathbb{P},r}(\mathcal{S}).$$

Finally, with probability at least $1 - 2\exp\left\{-n\delta^2 \mathrm{cut}_{\mathbb{P},r}(\mathcal{S})^2\right\}$,

$$(1-\delta)n(n-1)\mathrm{cut}_{\mathbb{P},r}(\mathcal{S}) \leq \mathrm{cut}_{n,r}(\mathcal{S}[X]) \leq (1+\delta)n(n-1)\mathrm{cut}_{\mathbb{P},r}(\mathcal{S}).$$

Proof of Lemma 4. The bounds on $d_{\min}$ and $d_{\max}$ follow from standard reasoning, in which we first use the regularity conditions (A) and (B) to upper and lower bound the expected degree $\mathbb{E}(\deg_{n,r}(x_i))$ over all x_i,

$$\lambda_{\min} \frac{\nu_d r^d}{a} \leq \min_{i=1,\dots,n} \mathbb{E}(\deg_{n,r}(x_i)) \leq \max_{i=1,\dots,n} \mathbb{E}(\deg_{n,r}(x_i)) \leq \lambda_{\max} \nu_d r^d,$$

and then apply Bernstein's inequality and a union bound (formally, Lemma) to control the maximal deviation $\deg_{n,r}(x_i) - \mathbb{E}(\deg_{n,r}(x_i))$. We state this second step as a separate Lemma, as it will be useful in contexts besides the proof of Lemma 4.

Hoeffding's inequality for binomial random variables leads to a bound on the deviation of $|S[X]|$ from its mean. Finally, the functionals $\mathrm{vol}(\mathcal{S}[X])$ and $\mathrm{cut}(\mathcal{S}[X])$ are U-statistics and therefore by Hoeffding's inequality concentrate around their respective expectations. □

Lemma 5 (Bernstein's inequality and a union bound.)**.** *For $M \geq 1$, let $\mathcal{A}_1, \dots, \mathcal{A}_M$ be subsets of $\mathbb{R}^d$, and denote the minimum probability $p_{\min} := \min_{m=1,\dots,M} \mathbb{P}(\mathcal{A}_m)$, and likewise let $p_{\max} := \max_{m=1,\dots,M} \mathbb{P}(\mathcal{A}_m)$. Then*

$$(1-\delta)p_{\min} \leq \min_{m=1,\dots,M} \mathbb{P}_n(M) \leq \max_{m=1,\dots,M} \mathbb{P}_n(M) \leq (1+\delta)p_{\max}$$

with probability at least $1 - 2M \exp\left\{ -\frac{\frac{1}{2}\delta^2 p_{\min} n}{1 + \frac{\delta}{3}} \right\}$.