Notes for week of 10/4/19 - 10/11/19

Alden Green

October 21, 2019

Let $G = (V, E)$ be an undirected, unweighted graph with $m = |E|$ total edges, defined on vertices $v_1, \ldots, v_n$. Let $A = (A_{ij})$ be the adjacency matrix of G, $W = (D^{-1}A + I)/2$ be the lazy random walk matrix over G, and for $s \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, define the PPR vector

$$p(\alpha, s) = \alpha s + (1 - \alpha) sp(\alpha, s)W.$$

(We will often drop the dependence on α, s and denote the PPR vector as p.)

Order the vertices $v_{(1)}, \ldots, v_{(n)}$ so that $p(v_{(1)})/d(v_{(1)}) \geq \ldots, p(v_{(n)})/d(v_{(n)})$, where $d(v_i) = \sum_{i=1}^n A_{ij}$ is the degree of v_i. We refer to

$$S_j = \left\{v_{(1)}, \ldots, v_{(j)}\right\}$$

as the jth sweep cut of the PPR vector.

It is known that if there exists a set $S \subset V$ with small normalized cut, then when the PPR vector is properly initialized, some sweep cut S_j of the PPR vector will itself have small normalized cut. Normalized cut is defined as

$$\Phi(S) = \frac{\sum_{v_i \in S} \sum_{v_j \in S^c} A_{ij}}{\min \text{vol}(S), 2m - \text{vol}(S)}$$

where $\text{vol}(S) = \sum_{v_i \in S} d(v_i)$ is the volume of S.

The proof that PPR is a good local partitioning algorithm (i.e. it concentrates on sets of small normalized cut) is heavily dependent on mixing time results for the PPR vector. We now give these results, before stating and proving some improvements.

1 Mixing of the PPR vector.

To quantify the mixing of the PPR vector, we introduce the function $p[\cdot] : [0, 2m] \to [0, 1]$. For $k_j = \text{vol}(S_j)$ for some sweep cut S_j, we let $p[k_j] = p(S_j)$, where we adopt the convention $p(S) = \sum_{v_i \in S} p_i$ for a vector $p \in \mathbb{R}^n$. We then define $p[\cdot]$ over its domain by interpolating between $0, p[k_1], \ldots, p[k_n]$. The mixedness of the PPR vector is then measured by the function $h : [0, 2m] \to [0, 1]$

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

1.1 Previous work.

The following theorem, given by Anderson,Chung,Lang, gives an upper bound on $h(k)$ as a function of $\Phi(p)$, where $\Phi(p) = \min_{j=1,\ldots,n} \Phi(S_j)$ is the smallest normalized cut of any sweep cut of p.

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

$$h(k) \leq \alpha t + \sqrt{\bar{k}} \left(1 - \frac{\phi^2}{8}\right)^t.$$

(where $\bar{k} = \min\{k, 2m - k\}$), or $\Phi(p) < \phi$.

The following result is essentially the contrapositive of Theorem 1.

Theorem 2. *If there exists a set $S \subset V$ and a constant $\delta \geq \frac{2}{\sqrt{m}}$ satisfying*

$$p(\alpha, s) - \frac{\mathrm{vol}(S)}{2m} > \delta,$$

then

$$\Phi(p) < \sqrt{\frac{18\alpha \log(m)}{\delta}}.$$

1.2 Improved bounds.

The presence of the factor of $\sqrt{\log(m)}$ is antagonistic to any work trying to prove consistency results, meaning results when $|V| \to \infty$. We therefore state and prove alternatives to Theorems 1 and 2, which allow us to avoid this factor of $\sqrt{\log(m)}$ under certain conditions. These theorems are related to the mixing time results proved by Lovasz,Simonovits. To state them, we must introduce some additional notation.For a given $0 \leq K_0 \leq m$, 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 $h(2m - K_0)$. Additionally, let

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

Theorem 3. *Let $p = p(\alpha, s)$ 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 $k \in [K_0, 2m - K_0]$:*

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

or $\Phi(p) < \phi$.

As a sanity check, we confirm that Theorem 3 is no weaker than Theorem 1. 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 1.

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

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

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

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

(Since h is not differentiable at points $k = \text{vol}(S_j)$, 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 $C(K_0)$. Let $d_{\min}$ and $d_{\max}$ be the minimum and maximum degrees of G, respectively.

Lemma 1. *Assume $s = \chi_v$ for some $v \in V$. If $S_1 = \{v\}$, let $K_0 = d(v)$ otherwise let $K_0 = 0$. Then,*

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

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

$$h'(2m - K_0) \geq \frac{d_{\max}}{d_{\min}\text{vol}(G)}. \tag{3}$$

As a result,

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

To bring Theorem 3 to bear, 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 $K_0 = \text{vol}(S_j)$ for some $j = 0, \ldots, n$. Then,*

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

Therefore, for any $k \in \mathbb{R}$,

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

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

Corollary 1. *Let $p = p(\alpha, \chi_v)$ 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 $k \in [d(v), 2m - d(v)]$:*

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

or $\Phi(p) < \phi$.

Proof. If $S_1 = \{v\}$, choose $K_0 = d(v)$; otherwise, choose $K_0 = 0$. As a result, Lemma 1 holds. Then, apply Theorem 3 and Lemma 2. □

It is worth briefly comparing Corollary 1 to Theorem 1. Corollary 1 will give better estimates of $h(k)$ when α is small, the degrees $d(u)$ are relatively uniform across $u \in V$, and $\frac{\sqrt{m}}{d_{\min}^2} << 1$, i.e. $d_{\min} >> \sqrt{n}$. This last point is the key: it reflects the fact that, all else being equal, random walks mix faster on graphs where all the very small sets (e.g singletons) have very large expansion. More nuanced improvements of this nature are discussed further in Kannan,Lovasz,Montenegro, albeit with respect to continuous space random walks, rather than the case of PPR over a graph we consider here.

We are now in a position to state the main result of this section. It is similar in form to Theorem 2, but reflects the improvements due to using Corollary 1 as opposed to Theorem 1.

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

$$p(S) - \frac{\mathrm{vol}(S)}{\mathrm{vol}(G)} > \delta \tag{4}$$

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 (4). 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 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 conductance of the PPR vector $\Phi(p) = \min_{j=1,\ldots,n} \Phi(S_j)$. First, we restate a theorem of ACL which lower bounds the probability mass $p(\alpha, s)(C)$ as a function of the normalized cut $\Phi(C)$.

Theorem 5. *For any set C and any constant α, there exists a subset $C_\alpha \subset C$ with $\mathrm{vol}(C_\alpha) \geq \mathrm{vol}(C)/2$, such that for any vertex $v \in C_\alpha$, the PPR vector $p(\alpha, \chi_v)$ satisfies*

$$p(\alpha, \chi_v)(C) \geq 1 - \frac{\Phi(C)}{\alpha}.$$

Combining Corollary 1 and Theorem 5 leads to an upper bound on $\Phi(p)$.

Theorem 6. *Let C be a set satisfying*

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

Suppose $10\Phi(C) \leq \alpha \leq 20\Phi(C)$. Then, there exists a subset $C_\alpha \subset C$ with $\mathrm{vol}(C_\alpha) \geq \mathrm{vol}(C)/2$ such that for any $v \in C_\alpha$,

$$\Phi(p(\alpha, \chi_v)) \leq \sqrt{3200\left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log 20\right\}\Phi(C)}$$

3 Local Partitioning of Neighborhood Graphs.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact domain, and let $\mathbb{P}$ be a probability distribution defined on $\mathcal{X}$ with density f. Sample $X = \{x_1, \ldots, x_n\}$ independently from f. For a given radius $r \in (0, \infty)$, the neighborhood graph $G := G_{n,r}$ has vertices $V = X$, and edges $E = \{(x_i, x_j) : |x_i - x_j| \leq r\}$.

To apply Theorem 6 when G is a random geometric graph, we will need estimates on the degree and volume graph functionals. These estimates will rely on certain regularity properties of $\mathcal{X}$ and 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 $a > 0$ such that for any $r > 0$ 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. Let $\mathcal{S} \subset \mathcal{X}$.

Lemma 3. *Let $X = \{x_1, \ldots, x_n\}$ be sampled independently from $\mathbb{P}$, and let $G := G_{n,r}$ be a neighborhood graph over X. Suppose $\mathcal{X}$ and f satisfy the regularity properties (A) and (B). Then, for any $\delta \in (0, 1)$ and any $r \in (0, \infty)$, there exists numbers $c_1 := c_1(f, \mathcal{X}, a)$ and $c_2 := c_2(f, \mathcal{X}, a)$ independent of the sample size n such that each of the following bounds hold with probability at least $1 - c_1 n \exp\{-c_2 n\}$:*

- $(1 - \delta)\lambda_{\min}\frac{\nu_d r^d}{a} n \leq d_{\min} \leq d_{\max} \leq (1 + \delta)\lambda_{\min}\frac{\nu_d r^d}{a} n$,
- $(1 - \delta)\mathbb{P}(\mathcal{S})n \leq |\mathcal{S}[X]| \leq (1 - \delta)\mathbb{P}(\mathcal{S})n$,
- $(1 - \delta)\mathbb{E}[\mathrm{vol}(\mathcal{S}[X])] \leq \mathrm{vol}(\mathcal{S}[X]) \leq (1 + \delta)\mathbb{E}[\mathrm{vol}(\mathcal{S}[X])]$,
- $(1 - \delta)\mathbb{E}[\mathrm{cut}(\mathcal{S}[X])] \leq \mathrm{cut}(\mathcal{S}[X]) \leq (1 + \delta)\mathbb{E}[\mathrm{cut}(\mathcal{S}[X])]$,

The bounds on $d_{\min}$ and $d_{\max}$ follow from standard reasoning, in which we i) use the regularity conditions (A) and (B) to upper and lower bound the expected degree $\mathbb{E}(d(x_i))$ over all x_i, ii) apply Bernstein's inequality to obtain bounds on the deviation $|d(x_i) - \mathbb{E}d(x_i)|$ for a given i, and iii) apply a union bound. Bernstein's inequality also 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 3 implies that when the sample size is sufficiently large, the conditions of Theorem 6 will be met. We therefore obtain the following upper bound on the normalized cut of PPR computed on a neighborhood graph.

Corollary 2. *Fix $\mathcal{S} \subset \mathcal{X}$, radius $r \in (0, \infty)$, and $\delta \in (0, 1/7]$. Suppose $\mathcal{X}$ and f satisfy the regularity properties (A) and (B), that*

$$n \geq a\frac{(1 + \delta)}{(1 - \delta)^2} \cdot \frac{\lambda_{\max}}{\lambda_{\min}} \max\left\{\frac{1}{\mathbb{P}(\mathcal{S})}, \frac{10a}{(1 - \delta)\lambda_{\min}\nu_d r^d}\right\}, \tag{5}$$

and

$$\mathbb{E}[\mathrm{vol}(\mathcal{S}[X])] \leq 2\frac{(1 - \delta)}{(1 + \delta)}\mathbb{E}[\mathrm{vol}((\mathcal{X} \setminus \mathcal{S})[X])], \tag{6}$$

and that $\Phi_{\mathbb{P}}(\mathcal{S}) \leq \frac{1}{200(1+\delta)}$. *Set* $\alpha = 15\Phi_{\mathbb{P}}(\mathcal{S})$. *Then, there exists a set* $\mathcal{S}[X]_\alpha \subset \mathcal{S}[X]$ *with* $\mathrm{vol}(\mathcal{S}[X]_\alpha) \leq \mathrm{vol}(\mathcal{S}[X])/2$, *such that for any* $v \in \mathcal{S}[X]_\alpha$,

$$\Phi(p(\alpha, \chi_v)) \leq \sqrt{3200\left\{2\log\left(\frac{a}{(1-\delta)\lambda_{\min}\nu_d r^d}\right) + \log 20\right\}\Phi(\mathcal{S}[X])}$$

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

The important takeaway is that $\Phi(p) \lesssim \sqrt{\Phi(\mathcal{S}[X])}$ (where we write $a_n \lesssim b_n$ when there exists constant c such that $a_n \leq c b_n$ for all n.) In particular, we have eliminated the factor of $\log(\mathrm{vol}(G))$ which made upper bounds derived from Theorem 2 facile as the sample size n and the volume $\mathrm{vol}(G) \to \infty$.

4 Implications for Density Clustering 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 $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$ be rectangles in $\mathbb{R}^2$,

$$\mathcal{C}_0 = \left[-\frac{3\sigma}{2}, -\frac{\sigma}{2}\right] \times \left[-\frac{\rho}{2}, \frac{\rho}{2}\right], \quad \mathcal{C}_1 = \mathcal{C}_0 + (\sigma, 0), \quad \mathcal{C}_2 = \mathcal{C}_0 + (2\sigma, 0)$$

and let $\mathbb{P}$ be the mixture distribution

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

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

$$f(x) = \frac{1}{\rho\sigma}\left(\frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_1) + \frac{1-\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_2) + \frac{\epsilon}{2}\mathbf{1}(x \in \mathcal{C}_0)\right)$$

so that for any $\epsilon < \lambda < (1-\epsilon)/2$, $\mathbb{C}_f(\lambda) = \{\mathcal{C}_1, \mathcal{C}_2\}$.

Sample $x_1, \ldots, x_n$ from f, and form the neighborhood graph G. Let $p(\alpha, \chi_v)$ be a PPR vector defined on G, and suppose our goal is to recover $\mathcal{C}_1$ using $\widehat{C} = \mathrm{argmin}_{i=1,\ldots,n}\Phi(S_j)$, the minimum conductance sweep cut of p. As the following Lemma demonstrates, even when $p(\alpha, \chi_v)$ is reasonably initialized, if the density cluster $\mathcal{C}_1$ is sufficiently geometrically ill-conditioned there are many seed nodes $v \in \mathcal{C}_1$ such that $\widehat{C}$ will fail to recover $\mathcal{C}_1$.

Theorem 7. *Let* $\alpha = 15\Phi_{\mathbb{P}}(\mathcal{L})$ *and* $r = \frac{1}{4}\sigma$. *Suppose* $\sigma < \frac{\rho}{1600}$. *Then, there exists a set* C_α *with* $|C_\alpha \cap \mathcal{C}_1[X]| \geq |\mathcal{C}_1[X]|/6$ *such that for any* $v \in C_\alpha$, *the minimum conductance sweep cut* $\widehat{C}$ *of* $p(\alpha, \chi_v)$ *satisfies*

$$\frac{\Delta(\widehat{C}, \mathcal{C}_1[X])}{n} \geq 1 - c\frac{1}{\epsilon^2}\sqrt{\log\left(\frac{\rho\sigma}{\epsilon\sigma^d}\right)\frac{\sigma}{\rho}}$$

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

Proof. Let $\mathcal{L}$ be the bottom half of the rectangle $\mathcal{X}$,

$$\mathcal{L} = \left[-\frac{3\sigma}{2}, \frac{3\sigma}{2}\right] \times \left[-\frac{\rho}{2}, \frac{\rho}{2}\right].$$

We will take for granted that with probability at least $1 - c_1 n \exp\{-c_2 n\}$, the following two statements are true

1. $\Phi_{n,r}(\mathcal{L}[X]) \leq 32\frac{r}{\rho}$, and
2. For any set $A \subset X$,

$$\Phi(A) \geq \frac{\epsilon^2 r}{\sigma}\left(1 - \frac{\left|\widehat{C}\triangle\mathcal{C}_1[X]\right|}{n(1-\epsilon)}\right) \tag{7}$$

It can be shown that $\mathcal{L}$ and $\mathcal{X}, f$ satisfy all of the conditions of Corollary 2. Therefore, choosing $\alpha = 10\Phi_{n,r}(\mathcal{L}[X])$, there exists a set C_α with $\mathrm{vol}(C_\alpha) \geq \mathrm{vol}(\mathcal{L}[X])/2$ such that

$$\Phi(\widehat{C}) \leq c\sqrt{\log\left(\frac{\rho\sigma}{\epsilon\sigma^d}\right)\Phi(\mathcal{L}[X])} \leq c\sqrt{\log\left(\frac{\rho\sigma}{\epsilon\sigma^d}\right)\frac{r}{\rho}}.$$

Combining this inequality with (7), we have

$$\frac{\epsilon^2 r}{\sigma}\left(1 - \frac{\left|\widehat{C}\triangle\mathcal{C}_1[X]\right|}{n(1-\epsilon)}\right) \leq c\sqrt{\log\left(\frac{\rho\sigma}{\epsilon\sigma^d}\right)\frac{r}{\rho}}.$$

Plugging in $r = \sigma/4$, and solving for $\frac{\left|\widehat{C}\triangle\mathcal{C}_1[X]\right|}{n}$, we obtain the desired result. □

5 Proofs.

5.1 Proof of Theorem 3.

The proof of Theorem 3 is essentially a combination of the proofs of Theorem 1 and Theorem 1.2 in Lovasz and Simonovits. We will show that if $\Phi(p) > \phi$, then (1) 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 $C(K_0)\cdot\sqrt{k} \geq h(k) - L_{K_0}(k)$ for all $k \in [K_0, 2m - K_0]$, which implies

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

Now, we proceed with the inductive step. We will show that (1) 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, it holds for all $k \in [K_0, 2m - K_0]$ by the concavity of the square root function.

By Lemma 4, we have

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

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

$$h(k_j) \leq \alpha + \frac{1}{2}\big(h(k_j - \phi\overline{k}_j) + h(k_j + \phi\overline{k}_j)\big) \tag{8}$$

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 (8), yielding

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

A Taylor expansion of $\sqrt{1+\phi}$ around $\phi = 0$ yields the following bound (see Lemma 5):

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

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 when valid, we have

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

5.2 Proof of Lemma 1.

The result of the Lemma is obvious once we show (2) and (3).

Assume $k < m$, and let $\mathrm{vol}(S_j) \leq k < \mathrm{vol}(S_{j+1})$ (where we let $S_0 = \emptyset$. The function h has the following representation:

$$h(k) = \sum_{i=0}^{j}\big(p(v_{(i)}) - \pi(v_{(i)})\big) + \frac{k - \mathrm{vol}(S_j)}{d(v_{(j+1)})}\big(p(v_{(j+1)}) - \pi(v_{(j+1)})\big) \tag{9}$$

where $\pi(s) = d(s)/\mathrm{vol}(G)$.

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

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

By Lemma 6, if $v_{(1)} = v$, then

$$p(\alpha, s)(v_{(2)}) \leq \frac{1}{2d_{\min}},$$

and if $v_{(1)} \neq v$ then the same inequality holds with respect to $p(\alpha, s)(v_{(1)})$. As a result, by (10), for either $K_0 = d(v)$ (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 (2).

Now assume $k \geq m$. The inequality (3) follows almost immediately from the representation (9), since

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

5.3 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 $h(2m - K_0)$,

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

To upper bound $h(K_0)$, we invoke the crude bounds of Lemma 6,

$$\begin{aligned} h(K_0) \leq p(S_j) &\leq p(\{v\}) + p(S_j \setminus \{v\}) \\ &\leq \frac{2\alpha}{1+\alpha} + \frac{|S_j|}{2d_{\min}} \\ &\leq \frac{2\alpha}{1+\alpha} + \frac{K_0}{2d_{\min}^2}, \end{aligned}$$

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

5.4 Proof of Theorem 6.

Since $\alpha \geq 10\Phi(C)$ and $v \in C_\alpha$, by Theorem 5,

$$p(\alpha, \chi_v)(C) \geq \frac{9}{10}.$$

This inequality along with the assumption $\mathrm{vol}(C) \leq \frac{2}{3}\mathrm{vol}(G)$ implies that $p(\alpha, \chi_v)(C) - \frac{\mathrm{vol}(C)}{\mathrm{vol}(G)} \geq \frac{1}{5}$. Since we assume $|C| \geq \frac{d_{\max}}{d_{\min}}$, the hypothesis of Theorem 4 is satisfied with $\delta = 1/5$, and we have

$$\Phi(p(\alpha, \chi_v)) \leq \sqrt{\frac{320\Phi(C)\left\{\log\left(\frac{m}{d_{\min}^2}\right) + \log\left(\frac{2}{\delta'}\right)\right\}}{\delta'}}$$

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.

5.5 Proof of Corollary 2.

Since we assume $\mathcal{X}$ and f satisfy the regularity conditions (A) and (B), the inequalities in Lemma 3 hold with probability at least $1 - nc_5 \exp\{-c_6 n\}$. We will condition on these inequalities throughout.

We now verify that the conditions required to apply Theorem 6 are met. The first two results in Lemma 3 imply that $\frac{d_{\max}}{d_{\min}} \lesssim |S[X]|$ and $\frac{d_{\max}}{d_{\min}^2} \lesssim 1$; the lower bound on n given in (5) ensures that these bounds hold exactly.

Since we additionally assume that $\Phi_{\mathbb{P}}(\mathcal{S}) \leq \frac{1-\delta}{200(1+\delta)}$, the third and fourth results in Lemma 3 imply the third condition in Theorem 6 is also met. Finally, setting $\alpha = 15\Phi_{\mathbb{P}}(\mathcal{S})$, these results imply that

$$10\Phi(S) \leq 15\frac{(1-\delta)}{1+\delta}\Phi(S) \leq \alpha \leq 15\frac{(1+\delta)}{1-\delta}\Phi(S) \leq 20\Phi(S)$$

where the first and last inequalities follow since $\delta \leq 1/7$. We may therefore apply Theorem 6 and obtain that

$$\Phi(p(\alpha, \chi_v)) \leq \sqrt{3200\left\{\log\left(\frac{\mathrm{vol}(G)}{d_{\min}^2}\right) + \log 20\right\}\Phi(C)}$$

Using the (potentially loose) bound $\mathrm{vol}(G) \leq n^2$, Corollary 2 follows from the first result in Lemma 3.

6 Technical results.

The following Lemma relates $p(S)$ to itself. It is Lemma 5 in Anderson,Chung,Lang.

Lemma 4. *For each $j \in [1, n-1]$,*

$$p[\mathrm{vol}(S_j)] \leq \alpha s(S_j) + \frac{1-\alpha}{2}\left(p[\mathrm{vol}(S_j) + |\partial(S_j)|] + p[\mathrm{vol}(S_j) - |\partial S_j|]\right)$$

where $\partial S = \{u \in S : \exists v \in S^c, (u,v) \in E\}$ is the boundary of S.

The following Lemma provides some Taylor expansions of the functions $\sqrt{1+x}$ and $\sqrt{1-x}$ about $x = 0$.

Lemma 5. *For any $x \in [0,1]$,*

$$\sqrt{1+x} \leq 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16}$$

and

$$\sqrt{1-x} \leq 1 - \frac{x}{2} - \frac{x^2}{8} - \frac{x^3}{16}.$$

Next, we provide some very crude bounds on the maximum value the PPR vector $p(\alpha, \chi_v)$ can attain.

Lemma 6. *For every $u \in V$,*

$$p(\alpha, \chi_v) \leq \mathbf{1}(u = v)\frac{2\alpha}{1+\alpha} + \frac{1}{2d_{\min}}.$$

Proof. 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

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

When $u = v$, we have that

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

and the result follows. □

6.1 Concentration Inequalities.

Bernstein's inequality. Let S_m be a binomial random variable, $S_m \sim \text{Bin}(n, p)$, and let $\mu = np$. Then, Bernstein's inequality gives

$$P\left(|S_m - \mu| \geq \delta\mu\right) \leq 2\exp\left\{-\frac{\frac{1}{2}\delta^2\mu}{1+\frac{\delta}{3}}\right\}$$

for any $\delta > 0$. Now suppose $S_{m1}, \ldots, S_{mm}$ are each binomial random variables, with $S_{mj} \sim \text{Bin}(n, p_j)$. Taking a union bound over $j = 1, \ldots, m$, we have that

$$P\left(\max_{j=1,\ldots,m} |S_m - \mu_j| \geq \delta\mu_j\right) \leq 2n\exp\left\{-\frac{\frac{1}{2}\delta^2\mu_{\min}}{1+\frac{\delta}{3}}\right\}$$

where $\mu_{\min} = \min_{j=1,\ldots,n} \mu_j$. Similarly letting $\mu_{\max} = \max_{j=1,\ldots,n} \mu_j$, we have that with probability at least $1 - 2n\exp\left\{-\frac{\frac{1}{2}\delta^2\mu_{\min}}{1+\frac{\delta}{3}}\right\}$,

$$(1-\delta)\mu_{\min} \leq S_{mj} \leq (1+\delta)\mu_{\max}, \quad \text{for all } j = 1, \ldots, m.$$

Hoeffding's Inequality for U-statistics. Let U_m be a degree-2 U-statistic with kernel h, i.e.

$$U_m = \sum_{i=1}^{m} \sum_{j \neq i} h(X_i, X_j).$$

where $X_1, \ldots, X_m$ are i.i.d random variables. Let $\mu_h = \mathbb{E}(U_m) = n(n-1)\mathbb{E}(h(X_1), h(X_2)) =: n(n-1)p_h$, and assume $\|h\|_\infty \leq 1$. Then, Hoeffding's inequality gives

$$P(|U_n - \mu| \geq \delta\mu) \leq 2\exp\left(-\frac{\delta^2\mu_h p_h}{4}\right).$$

Therefore, with probability at least $1 - 2\exp\left(-\frac{\delta^2\mu_h p_h}{4}\right)$,

$$(1-\delta)\mu_h \leq U_n \leq (1+\delta)\mu_h.$$

7 Old Stuff.

7.1 Proof of Lemma 3.

Each statement in Lemma 3 is a standard consequence of the concentration results of 6.1. For each $i = 1, \ldots, n$, conditional on x_i the degree functional $d(x_i)$ follows a binomial distribution $d(x_i) \sim \text{Bin}(n -$

$1, \mathbb{P}(B(x_i, r)))$. Moreover by the regularity properties, for any $x \in \mathcal{X}$,

$$\lambda_{\min} \frac{\nu_d r^d}{a} \leq \mathbb{P}(B(x, r)) \leq \lambda_{\max} \nu_d r^d.$$

Therefore, for any $\delta > 0$, applying Bernstein's inequality and a union bound yields

$$\lambda_{\min}(1-\delta)\frac{\nu_d r^d}{a} n \leq d_{\min} \leq d_{\max} \leq \lambda_{\max}(1-\delta)\nu_d r^d n \tag{11}$$

with probability at least $1 - 2n \exp\left\{-\frac{\delta^2 n \lambda_{\min} \nu_d r^d}{a\left(1+\frac{\delta}{3}\right)}\right\}$. Additionally $|S[X]| \sim \mathrm{Bin}(n, \mathbb{P}(S))$, and therefore

$$(1-\delta)\mathbb{P}(\mathcal{S})n \leq |S[X]| \tag{12}$$

with probability at least $1 - \exp\left\{-\frac{\delta^2 n \mathbb{P}(\mathcal{S})}{1+\frac{\delta}{3}}\right\}$.

The inequalities (11) and (12) imply that with high probability $\frac{d_{\max}}{d_{\min}} \lesssim |S[X]|$ and $\frac{d_{\max}}{d_{\min}^2} \lesssim 1$; the lower bound on n given in (5) ensures that these bounds hold exactly.

Finally, the statement $\mathrm{vol}\big(\mathcal{S}[X]\big) \leq \frac{2}{3}\mathrm{vol}(G)$ is equivalent to $\mathrm{vol}\big(\mathcal{X} \setminus \mathcal{S}[X]\big) \geq \frac{1}{2}\mathrm{vol}\big(\mathcal{S}[X]\big)$. For any set $\mathcal{A} \subseteq \mathbb{R}^d$, the volume $\mathrm{vol}\big(\mathcal{A}[X]\big)$ is a U-statistic, since it can be written as

$$\mathrm{vol}\big(\mathcal{A}[X]\big) = \sum_{i=1}^{m} \sum_{j \neq i} \underbrace{\frac{\mathbf{1}(x_i \in \mathcal{S}) \vee \mathbf{1}(x_j \in \mathcal{S})}{2} \mathbf{1}(\|x_i - x_j\| \leq r)}_{h_{\mathcal{A}}(x_i, x_j)}$$

Let $p_{h_{\mathcal{A}}} = \mathbb{E}[h_{\mathcal{A}}(X_1, X_2)]$ for X_1 and X_2 independent random variables with distribution $\mathbb{P}$. By Hoeffding's inequality for U-statistics we have

$$\mathrm{vol}\big(\mathcal{S}[X]\big) \leq (1+\delta)\mathbb{E}\left[\mathrm{vol}(\mathcal{S}[X])\right], \quad \mathrm{vol}\big(\mathcal{S}[X]\big) \leq (1-\delta)\mathbb{E}\left[\mathrm{vol}(\mathcal{X} \setminus \mathcal{S}[X])\right] \tag{13}$$

with probability at least $1 - 2\exp\left\{-\frac{\delta^2 (n-1)^2 \min\{p_{h_{\mathcal{S}}}, p_{h_{\mathcal{X}\setminus\mathcal{S}}}\}}{4}\right\}$. If (13) is satisfied, then by (6) the inequality $\mathrm{vol}\big(\mathcal{X} \setminus \mathcal{S}[X]\big) \geq \frac{1}{2}\mathrm{vol}\big(\mathcal{S}[X]\big)$ holds, and therefore $\mathrm{vol}\big(\mathcal{S}[X]\big) \leq \frac{2}{3}\mathrm{vol}(G)$.