Local Spectral Clustering of Density Upper Level Sets

Anonymous Authors[1]

Abstract

Spectral clustering methods are a family of popular non-parametric clustering tools. Recent works have proposed and analyzed *local* spectral methods, which extract clusters using locally-biased random walks around a user-specified seed node, and are known to have worst-case guarantees for certain graph-based measures of cluster quality. In contrast to existing works, we analyze the personalized PageRank (PPR) algorithm in the classical statistical learning setup, where we obtain samples from an unknown distribution, and aim to identify connected regions of high-density (density clusters). We introduce a bicriteria for evaluating cluster quality and provide guarantees for these criteria evaluated on empirical analogues to these density clusters. As a result, the PPR algorithm run over a neighborhood graph is shown to extract sufficiently salient density clusters.

1. Introduction

Let $\mathbf{X} := (x_1, \dots, x_n)$ be a sample drawn i.i.d from a distribution $\mathbb{P}$ with density f supported on $\mathbb{R}^d$. Our statistical learning task is clustering: splitting data into groups which satisfy some notion of within-group similarity and between-group difference.

Spectral clustering methods are a family of powerful non-parametric clustering algorithms. Let $G = (V, E)$ be an unweighted and undirected graph, with $\mathbf{A} \in \mathbb{R}^{V \times V}$ the symmetric adjacency where $A_{uv} = 1$ if $(u, v) \in G$. We form the random walk transition probability matrix $\mathbf{W}$, and corresponding graph Laplacian matrix $\mathbf{L}$[1]:

$$\mathbf{W} := \mathbf{D}^{-1}\mathbf{A}; \quad \mathbf{L} = \mathbf{I}_n - \mathbf{W} \tag{1}$$

where the degree matrix $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{uu} := \sum_{v \in V} \mathbf{A}_{uv}$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Roughly speaking, spectral clustering techniques first learn an embedding the data $\mathbf{X}$ using the spectrum of the graph Laplacian, and subsequently apply a simple clustering technique (such as k-means) to this *spectral embedding*. When applied to large graphs (or large point clouds) classical global spectral methods can be computationally cumbersome and insensitive to the local geometry of the distribution of the samples (Mahoney et al., 2012; Leskovec et al., 2010). This in turn has led to the investigation of local spectral algorithms (Spielman & Teng, 2013; Andersen et al., 2006; Leskovec et al., 2010) which leverage locally-biased spectra computed using random walks around a user-specified seed node.

We are interested in understanding what the output of a clustering algorithm on $\mathbf{X}$ reveals about the unknown density f. For $\lambda > 0$ and the upper level set $\{x : f(x) \geq \lambda\}$, it is intuitive (Hartigan, 1981; Chaudhuri & Dasgupta, 2010) to define clusters as the connected components $\mathbb{C}_f(\lambda)$ of the upper level set; we call these connected regions of high density *density clusters*, and study the ability of spectral methods to identify such clusters.

Graph connectivity criteria. A somewhat more standard mode of understanding spectral clustering methods is to view them as approximating some graph connectivity criteria. There are many graph-based measures which assess the cluster quality of a subset $S \subseteq V$ (or more generally the quality of a partition $S_1 \cup \ldots \cup S_m = V$, for $m \geq 2$. See (Yang & Leskovec, 2015; Fortunato, 2010) for an overview.)

Arguably a natural way to assess cluster quality is via a pair of criteria capturing the *external* and *internal connectivity* of S, respectively. We define the external connectivity of a subset $S \subseteq V$ to be its normalized cut. For $S' \subset V$, the cut between S and S' is

$$\mathrm{cut}(S, S'; G) := \sum_{u \in S} \sum_{v \in S'} \mathbf{1}\left((u, v) \in E\right)$$

and the volume of S is

$$\mathrm{vol}(S; G) := \sum_{u \in S} \sum_{v \in V} \mathbf{1}\left((u, v) \in E\right).$$

Then, the *normalized cut* of S is

$$\Phi(S; G) := \frac{\mathrm{cut}(S; G)}{\min\left\{\mathrm{vol}(S; G), \mathrm{vol}(S^c; G)\right\}} \tag{2}$$

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Often, either of the Laplacian matrices $\mathbf{L}_{sym} := \mathbf{D}^{\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ or $\mathbf{L}_{unn} := \mathbf{D} - \mathbf{A}$ are used instead.

(where we abbreviate $\mathrm{cut}(S;G) = \mathrm{cut}(S;S^c;G)$).

Given $S \subseteq V$, the subgraph induced by S is given by $G[S] = (S, E_S)$, where $(u,v) \in E_S$ if both u and v are in S and $(u,v) \in E_S$. Our internal connectivity parameter $\Psi(S)$ will capture the time it takes for the random walk over $G[S]$ to mix (that is, approach a stationary distribution) uniformly over S. Letting $|S| = m$, $\mathbf{A}_S$ denotes the $m \times m$ adjacency matrix representation of $G[S]$; similarly, $\mathbf{D}_S$ is the diagonal degree matrix with entries $(\mathbf{D}_S)_{ii} = \sum_{j:v_j \in S} \mathbf{A}_{ij}$, and $\mathbf{W} = \mathbf{D}_S^{-1}\mathbf{A}_S$ is the corresponding random walk matrix (again, defined only over $G[S]$.)

For $v \in V$ we write

$$\boldsymbol{\pi} = (\pi_u)_{u \in S}, \;\; q_{vu}^{(m)} = e_v \mathbf{W}_S^m e_u \tag{3}$$

for the stationary distribution and the m-step transition probability of the random walk over $G[S]$ originating at v. [2] (Of course, given the definition of $\mathbf{W}_S$ it is of course well known that the stationary distribution $\widetilde{\boldsymbol{\pi}}$ is given by $\widetilde{\pi}_u = (\mathbf{D}_S)_{uu}/\mathrm{vol}(S; G[S])$.)

The *relative pointwise mixing time* is defined as

$$\tau_\infty(\mathbf{q}; G[S]) := \min\left\{ m : \forall u,v \in V, \frac{\left| q_{vu}^{(m)} - \pi_u \right|}{\pi_u} \le 1/4 \right\}$$

where $\mathbf{q} = (\mathbf{q}_v^{(1)}, \mathbf{q}_v^{(2)}, ...)_{v \in V}, \mathbf{q}_v^{(m)} = (q_{vu}^{(m)})_{u \in V}$.

The internal connectivity parameter $\Psi(S;G)$ is simply one over the mixing time:

$$\Psi(S;G) = \frac{1}{\tau_\infty(\mathbf{q}; G[S])} \tag{4}$$

The graph clustering task is thus find a subset S (or, for global algorithms, a partition $S_1 \cup \ldots \cup S_m = V$), which has both small external and large internal connectivity. If S has normalized cut no greater than Φ, and inverse mixing time no less than Ψ, we will refer to it as a (Φ, Ψ)-cluster. Both local (Zhu et al., 2013) and global (Kannan et al., 2004) spectral algorithms have been shown to output clusters (or partitions) which provably satisfy approximations to the optimal (Φ,Ψ)-cluster (or partition) for a given graph G. [3]

Personalized PageRank. As mentioned previously, global algorithms which find spectral cuts may be computationally infeasible for large graphs; in this setting, local algorithms may be preferred or even required. We will restrict our attention in particular to one such popular algorithm: *personalized PageRank* (PPR). The personalized PageRank algorithm was first introduced by (Haveliwala, 2003) and variants of this algorithm have been studied further in several recent works (Spielman & Teng, 2011; 2014; Zhu et al., 2013; Andersen et al., 2006; Mahoney et al., 2012).

The random walk matrix $\mathbf{W}$ over the graph $G = (V,E)$ with associated adjacency matrix $\mathbf{A}$ is given by (1). PPR is then defined with respect to the following inputs: a user-specified seed node $v \in V$, and $\alpha \in [0,1]$ a teleportation parameter. Letting e_v be the indicator vector for v (meaning e_v has a 1 in the vth location and 0 everywhere else), the *PPR vector* is the solution to the following linear equation:

$$\mathbf{p}(v,\alpha;G) := \alpha \mathbf{e}_v + (1-\alpha)\mathbf{p}(v,\alpha;G)\mathbf{W} \tag{5}$$

We note in passing that, for $\alpha > 0$, the vector $\mathbf{p}(v,\alpha;G)$ can be well-approximated by a simple local computation (of a random walk with restarts at the node v.) We also point out that, from a density clustering standpoint, since density clusters are inherently local, using the PPR algorithm eases the analysis, and as we will observe in the sequel our analysis requires fewer global regularity conditions relative to more classical global spectral algorithms.

To compute a cluster $\widehat{C} \subset V$ using the PPR vector, we will take sweep cuts of $\mathbf{p}(v,\alpha;G)$. To do so we require another parameter π_0. For a number $\beta \ge 0$, the sweep cut S_β is then

$$S_\beta = \{u \in V : p_u > \beta\pi_0\}\,. \tag{6}$$

We will choose one such sweep cut to be our cluster estimate $\widehat{C}$.

Neighborhood graph. To formally introduce the local clustering algorithm we consider, we must first give a method for forming a graph over the data $\mathbf{X}$.

Let (r_n) be a sequence of positive numbers. Given a sequence of kernel functions $k_n : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ of the form $k_n(x,x') = k(\|x - x'\|/r_n)$ for k a non-increasing function, and data $\mathbf{X} = \{x_1, \ldots, x_n\}$ sampled from $\mathbb{P}$ as before, form the *neighborhood graph* $G_n = (\mathbf{X}, E_n)$ with $E_n = \{k(x_i,x_j) : 1 \le i < j \le n\}$. (Here, $\|\cdot\|$ is used to denote Euclidean norm).

Algorithm 1 will be the simple variant of PPR we analyze. It will take as input the data $\mathbf{X}$ along with user-specified parameters r, α, π_0, and $v \in \mathbf{X}$.

[2] Given a starting node v and and a random walk defined by transition probability matrix $\mathbf{P}$, the rotation $e_v\mathbf{P}^t$ is used to denote the distribution of the random walk after t steps.

[3] In the case of (Kannan et al., 2004), the internal connectivity parameter ϕ is actually the conductance, i.e. the minimum normalized cut within the subgraph $G[S]$. See Theorem 3.1 for details; however, note that $\phi^2/\log(\mathrm{vol}(S)) \le O(\Psi)$, and so the lower bound on ϕ translates to a lower bound on Ψ.

```
Algorithm 1 PPR on a neighborhood graph
Input: data $\mathbf{X}$, radius $r$, teleportation parameter $\alpha \in [0,1]$, seed node $v \in \mathbf{X}$, target $\pi_0$.
Output: $\widehat{C} \subset V$.
  1: Form the neighborhood graph $G_{n,r}$ as given in (8)
  2: Compute PPR vector $\mathbf{p}(v, \alpha; G_{n,r})$ as defined by (5).
  3: For $\beta \in (\frac{3}{10}, \frac{1}{2})$ compute sweep cuts $S_\beta$ as defined by (6).
  4: Return
       $$\widehat{C} = \arg\min_{\beta \in (\frac{3}{10}, \frac{1}{2})} \Phi(S_\beta; G_{n,r})$$
```

We note that the choice of a specific interval such as $(\frac{3}{10}, \frac{1}{2})$ for the range of β is standard in the analysis of PPR algorithms.

It is worth calling attention to some other work on computing the normalized cut over neighborhood graphs. In this context, continuous analogues to (for instance) normalized cut have been defined, over the data-manifold rather than the graph, and convergence finite sample graph-theoretic functionals to their continuous counterparts has been shown (Trillos et al., 2016; Ery et al., 2012; Maier et al., 2011). However, in addition to the graph-minimization problem being computationally infeasible, these continuous analogues are not always easily interpretable – and their corresponding minimizers not always easily identifiable – for the particular density function under consideration. Of course, relating these partitions to the arguably more simply defined high density clusters can be also challenging in general. Intuitively, however, under the right conditions such high-density clusters should have more edges within themselves than to the remainder of the graph. We formalize this intuition next.

1.1. Summary of results

Hereafter, we consider the uniform kernel function for a fixed $r > 0$,

$$k(x, x') = \mathbf{1}(\|x - x'\| \leq r) \tag{7}$$

and the associated neighborhood graph

$$G_{n,r} = (\mathbf{X}, E_{n,r}),\ (x_i, x_j) \in E_{n,r} \text{ if } k(x_i, x_j) = 1 \tag{8}$$

For a given high density cluster $\mathcal{C} \subseteq \mathbb{C}_f(\lambda)$, we call $\mathcal{C}[\mathbf{X}] = \mathcal{C} \cap \mathbf{X}$ the *empirical density cluster*. We now introduce a notion of consistency for the task of density cluster estimation:

Definition 1 (Consistent density cluster estimation)**.** *For an estimator $\widehat{\mathcal{C}}_n \subset \mathbf{X}$, and any $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_f(\lambda)$, we say $\widehat{\mathcal{C}}_n$ is a consistent estimator of $\mathcal{C}$ if the following statement holds: as the sample size $n \to \infty$, each of the following*

$$\mathcal{C}[\mathbf{X}] \subseteq \widehat{\mathcal{C}}_n, \text{ and } \widehat{\mathcal{C}}_n \cap \mathcal{C}'[\mathbf{X}] = \emptyset \tag{9}$$

occur with probability tending to 1.

Our results can now be summarized by the following points:

1. Under a natural set of geometric conditions [4], the normalized cut of an empirical density cluster $C[\mathbf{X}]$ can be bounded. Theorem 1 provides an upper bound.

2. Under largely the same set of geometric conditions to Theorem 1, Theorem 2 provides a lower bound on the inverse mixing time of a random walk over $C[\mathbf{X}]$. Theorem 3 gives a tighter lower bound, but requires the additional (restrictive) assumption of convexity.

3. In Section 3, we show these bounds on the graph connectivity criteria have algorithmic consequences for personalized PageRank. we show that a careful analysis of the form typical to local clustering algorithms yields Theorem 4, which states that Algorithm 1, properly initialized, performs consistent density cluster estimation in the sense of (9).

4. Corollary 1 follows as an immediate consequence of Theorems 1 and 3, along with the previous work of (Zhu et al., 2013). It gives an upper bound on the normalized cut of the set $\widehat{C}$ output by Algorithm 1, as well as upper bounding the symmetric set difference between $\widehat{C}$ and $\mathcal{C}[\mathbf{X}]$.

Organization. In Section 4, we provide some example density functions, to clarify the relevance of our results, and empirically demonstrate that violations of the geometric conditions we set out in Section 2 manifestly impact density cluster recovery (i.e. the conditions are not superfluous), before concluding in 5. First, however, we summarize some related work.

1.2. Related Work

In addition to the background given above, a few related lines of work are worth highlighting.

Global spectral clustering methods were first developed in the context of graph partitioning (Fiedler, 1973; Donath & Hoffman, 1973) and their performance is well-understood in this context (see, for instance, (Tolliver & Miller, 2006; von Luxburg, 2007)). In a similar vein, several recent works (McSherry, 2001; Lei & Rinaldo, 2015; Rohe et al., 2011;

[4] We formally introduce the geometric conditions in Section 2. They preclude clusters which are too thin and long, or those for which the gap in density between the high density area and the outside is not sufficiently large

Abbe, 2018; Chaudhuri et al., 2012; Balakrishnan et al., 2011) have studied the efficacy of spectral methods in successfully recovering the community structure in various variants of the stochastic block model.

Building on the work of Koltchinskii & Gine (2000) the works (von Luxburg et al., 2008; Hein et al., 2005) for instance, have studied the limiting behaviour of spectral clustering algorithms. These works show that when samples are obtained from a distribution, following appropriate graph construction, in certain cases the spectrum of the Laplacian converges to that of the Laplace-Beltrami operator on the data-manifold. However, relating the partition obtained using the Laplace-Beltrami operator, to the more intuitively defined high-density clusters, can be challenging in general.

Perhaps most similar to our results are (Vempala & Wang, 2004; Shi et al., 2009; Schiebinger et al., 2015), which study the consistency of spectral algorithms in recovering the latent labels in certain parametric and non-parametric mixture models. These results focus on global rather than local algorithms, and as such impose global rather than local conditions on the nature of the density. Moreover, they do not in general ensure recovery of density clusters which is a focus of our work.

2. Graph Quality Criteria on Well-Conditioned Density Clusters.

In order to provide meaningful bounds on the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$, we must introduce some assumptions on the density f.

Let $B(x,r) = \bigl\{y \in \mathbb{R}^d : \|y - x\| \leq r\bigr\}$ be a closed ball of radius r around the point x. Given a set $\mathcal{A} \subset \mathbb{R}^d$, and a number $\sigma > 0$, define the σ-expansion of $\mathcal{A}$ to be $\mathcal{A}_\sigma = \mathcal{A} + B(0,\sigma) = \bigl\{y \in \mathbb{R}^d : \inf_{x \in \mathcal{A}} \|y - x\| \leq \sigma\bigr\}$. We are now ready to give the assumptions, which we state with respect to a density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$ for some $\lambda > 0$, and expansion parameter $\sigma > 0$:

(A1) *Cluster separation:* For all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$,

$$\mathrm{dist}(\mathcal{C}_\sigma, \mathcal{C}'_\sigma) > \sigma,$$

where $\mathrm{dist}(\mathcal{A}, \mathcal{A}') = \min_{x \in \mathcal{A}} \mathrm{dist}(x, \mathcal{A}')$ for $\mathcal{A}' \subset \mathbb{R}^d$.

(A2) *Cluster diameter:* There exists $D < \infty$ such that for all $x, x' \in \mathcal{C}_\sigma$:

$$\|x - x'\| \leq D.$$

(A3) *Bounded density within cluster:* There exist numbers $0 < \lambda_\sigma < \Lambda_\sigma < \infty$ such that:

$$\lambda_\sigma = \inf_{x \in \mathcal{C}_\sigma} f(x) \leq \sup_{x \in \mathcal{C}_\sigma} f(x) \leq \Lambda_\sigma \qquad (10)$$

(A4) *Low noise density:* For some $\gamma > 0$, there exists a constant $c_1 > 0$ such that for all $x \in \mathbb{R}^d$ with $0 < \mathrm{dist}(x, \mathcal{C}_\sigma) \leq \sigma$,

$$\inf_{x' \in \mathcal{C}_\sigma} f(x') - f(x) \geq c_1 \mathrm{dist}(x, \mathcal{C}_\sigma)^\gamma,$$

where $\mathrm{dist}(x, \mathcal{A}) = \min_{x_0 \in \mathcal{A}} \|x - x_0\|$ for $\mathcal{A} \subset \mathbb{R}^d$.

We note that σ plays several roles here, precluding arbitrarily narrow clusters and long clusters in (A2) and (A3), flat densities around the level set in (A4), and poorly separated clusters in (A1).

Assumptions (A1), (A3) and (A4) are used to upper bound $\Phi(\mathcal{C}[\mathbf{X}]; G_{n,r})$, whereas (A2) and (A3) are necessary to lower bound $\Psi(\mathcal{C}[\mathbf{X}]; G_{n,r})$. We note that the lower bound on minimum density in (10) and (A1) combined are similar to the (σ, ϵ)-saliency of (Chaudhuri & Dasgupta, 2010), a standard density clustering assumption, while (A4) is seen in, for instance, (Singh et al., 2009), (as well as many other works on density clustering and level set estimation.) It is worth highlighting that these assumptions are all local in nature, a benefit of studying a local algorithm such as PPR.

We are ready to provide bounds on the graph quality bicriteria. For notational simplicity, hereafter for $S \subseteq \mathbf{X}$ we will refer to $\Phi(S; G_{n,r})$ as $\Phi_{n,r}(S)$, and likewise with $\Psi(S; G_{n,r})$ and $\Psi_{n,r}(S)$. We will also use $\nu(\cdot)$ to denote the uniform measure over $\mathbb{R}^d$, and $\nu_d = \nu\bigl(B(0,d)\bigr)$ as the measure of the unit ball.

We begin with an upper bound on the normalized cut in Theorem 1. We will require Assumptions (A1), (A3) and (A4) to hold; however, the upper bound on density in (10) will not be needed and so we omit the parameter Λ_σ from the statement of the theorem.

Theorem 1. *For some $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1), (A3) and (A4) for some $\sigma, \lambda_\sigma, c_1, \gamma > 0$. Then, for any $r < \sigma/4d$ and $\delta \in (0,1]$, the following statements hold with probability at least $1 - \delta$: Fix $\epsilon > 0$. Then, for*

$$n \geq \frac{9\log(2/\delta)}{\epsilon^2} \left(\frac{1}{\lambda_\sigma^2 \nu(\mathcal{C}_\sigma) \nu_d r^d}\right)^2 \qquad (11)$$

we have

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])}{r} \leq 4 c_\sigma d \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon \qquad (12)$$

where $c_\sigma = 1/\sigma$.

Remark 1. The proof of Theorem 1, along with all other proofs, can be found in the supplementary document. The key point is to note that for any $x \in \mathcal{C}$, the simple, (possibly loose) $B(x,\sigma) \subset \mathcal{C}_\sigma$ translates to the upper bound

$\nu(\mathcal{C}_{\sigma+r}) \leq (1 + 2dr/\sigma)\nu(\mathcal{C}_\sigma)$. We leverage (A4) to find a corresponding bound on the weighted volume, before applying standard concentration inequalities to convert from population to sample based results.

Remark 2. (12) is almost tight. Specifically, choosing

$$\mathcal{A}_\sigma = B(0,\sigma),$$
$$f(x) = \begin{cases} \lambda \text{ for } x \in \mathcal{A}_\sigma, \\ \lambda - \operatorname{dist}(x, \mathcal{A}_\sigma)^\gamma \text{ for } 0 < \operatorname{dist}(x,\mathcal{A}_\sigma) < r \end{cases}$$

we have that for n within constant order of the lower bound in (11), with probability at least $1 - \delta$

$$\frac{\Phi_{n,r}(\mathcal{A}_\sigma[\mathbf{X}])}{r} \geq c\frac{(\lambda - \frac{r^{\epsilon+1}}{\epsilon+1})}{\lambda} - \epsilon$$

for some constant c. (Note that a factor of $1/\sigma$ in c_σ is not replicated in this lower bound.)

We now provide a lower bound on $\Psi_{n,r}(\mathcal{C}[\mathbf{X}])$.

Theorem 2. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A2) and (A3) for some $\sigma, \lambda_\sigma, \Lambda_\sigma, D > 0$. Then, for any $r < \sigma/4d$, the following statement holds with probability at least $1-\delta$: Fix $0 < \epsilon > 1$. Then, for n satisfying:*

$$\sqrt{3^{d+1}\frac{(\log n + d\log\mu + \log(4/\delta))}{n\nu_d r^d \lambda_\sigma}} \leq \epsilon$$

we have

$$1/\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) \leq (d\log\mu + c_\lambda)\cdot \left(c_1 + c_2\frac{\Lambda_\sigma^4 D^{2d}}{\lambda_\sigma^4 r^{2d}}(d\log\mu + c_\lambda)\right) \quad (13)$$

where $\mu = \log(\frac{2D}{r})$; c_1, c_2, and c_3 are constants which depend only on dimension; and $c_\lambda = \log(\Lambda_\sigma^2/\lambda_\sigma^2)$.

Remark 3. The proof of Theorem 2 relies on upper bounding the mixing time using the *conductance* of $G_{n,r}[\mathcal{C}_\sigma[X]]$,

$$\widetilde{\Phi} = \min_{S \subset \mathcal{C}_\sigma[\mathbf{X}]} \Phi(S; G_{n,r}[\mathcal{C}_\sigma[X]])$$

The factor of $\frac{1}{r^{2d}}$ present in the bound of (13) is suboptimal. This exponential dependence on d stems from a loose bound on the aforementioned conductance in the proof of Theorem 2.In particular, we assert only that any set $S \subset \mathcal{C}_\sigma[X]$ must have $\operatorname{cut}(S; \mathcal{C}_\sigma[\mathbf{X}])$ on the order of $n^2 r^{2d}$, while upper bounding $\operatorname{vol}(S; \mathcal{C}_\sigma[\mathbf{X}])$ by roughly $n^2 r^d$, for a bound on the conductance of order r^d. The presence of r^{2d} comes from upper bounding the mixing time by about $1/\widetilde{\Phi}^2$, this latter bound being a variant of classic results on rapid mixing (Jerrum & Sinclair, 1989).

It is possible to sharpen the dependency on d, but at the cost of an additional assumption:

(A5) $\mathcal{C}$ is convex.

Additionally, the resulting bound holds only asymptotically.

Theorem 3. *Fix $\lambda > 0$, and let the conditions of Theorem 2 hold with respect to $\mathcal{C} \in \mathbb{C}_f(\lambda)$. Additionally, let $\mathcal{C}$ satisfy (A5). Then, for any $r < \sigma/4d$, the following statement holds: with probability one*

$$\liminf_{n\to\infty}\frac{1}{\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])} \geq c_1\frac{\Lambda_\sigma^8}{\lambda_\sigma^8}(\log\mu + c_\lambda)\Bigg(c_\lambda(d^3\log\mu + c_2) + \frac{d^2D^2}{r^2}\Big(\log d + c_2\frac{\Lambda_\sigma^2}{\lambda_\sigma^2}c_\lambda + c_3 + \log\log\mu\Big)\Bigg) + c_d\frac{o_r(1)}{r^2}, \quad (14)$$

where $\mu = \log(\frac{2D}{r})$, c_1, c_2 and c_3 are all global constants, c_d is constant in r but may depend on other quantities, and $c_\lambda = \log(\Lambda_\sigma^2/\lambda_\sigma^2)$. Additionally, $\lim_{r\to 0} o_r(1) = 0$.

We achieve superior rates in Theorem 3 to those of Theorem 2 in part by working with a generalization of the conductance, the *conductance function*

$$\widetilde{\Phi}_n(t) = \min_{\substack{S\subset\mathcal{C}_\sigma[\mathbf{X}] \\ \pi(S)\leq t}} \Phi(S; G_{n,r}[\mathcal{C}_\sigma[X]])$$

where $\pi(S)$ is the stationary distribution over $G_{n,r}[\mathcal{C}_\sigma[\mathbf{X}]]$. The utility of the conductance function comes from the known upper bound of mixing time by

$$\int \frac{1}{t\widetilde{\Phi}_n(t)^2}dt \quad (15)$$

which results in a tighter bound than merely using the conductance when $\widetilde{\Phi}_n(t)$ is large for small values of t.

Remark 4. The only potentially exponential dependence on dimension comes from the factor of c_d. However, for sufficiently small values of r this will be dominated by the preceding factors, due to the presence of the $o_r(1)$ term.

Remark 5. The proof of Theorem 3 relies on a to the best of our knowledge novel uniform lower bound on this conductance function over $G_{n,r}$ in terms of a population-level analogue, which we term $\widetilde{\Phi}_{\mathbb{P},r}$. Plugging $\widetilde{\Phi}_{\mathbb{P},r}$ into (15) yields a bound on mixing time (with respect to total variation distance) of order $dD^2/r^2 + d^2\log(D/r)$ over convex sets. By contrast, (14) is of order $d^2D^2/r^2 + d^3\log(D/r)$ (ignoring the $o_r(1)$ term) but handles mixing time with respect to relative pointwise distance (which is known to be a stricter metric.)

Remark 6. The convexity requirement is necessary only to lower bound the $\widetilde{\Phi}_{\mathbb{P},r}$; any bound on $\widetilde{\Phi}_{\mathbb{P},r}$ which does not require convexity could immediately be plugged into the machinery of the proof of Theorem 3 to achieve a corresponding result. Recent work (Abbasi-Yadkori et al., 2017) has developed such population-level bounds on the conductance function, however the Markov chain dealt with there is somewhat different than the one we consider.

3. Consistent cluster estimation with PPR.

For PPR to successfully recover density clusters, the ratio $\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])/\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])$ must be small.

We introduce

$$\begin{aligned}\mathbf{\Phi}(\sigma,\lambda,\lambda_\sigma,\gamma) &:= \frac{\lambda}{\lambda_\sigma}\frac{(\lambda_\sigma - \frac{\sigma^\gamma}{\gamma+1})}{\lambda_\sigma} \\ 1/\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D) &:= (d\log\mu + c_\lambda)\cdot \\ &\left(c_1 + c_2\frac{\Lambda_\sigma^4 D^{2d}}{\lambda_\sigma^4 r^{2d}}\left(d\log\mu + c_\lambda\right)\right)\end{aligned}$$

Well-conditioned density clusters satisfy all of the given assumptions, for parameters which results in 'good' values of $\mathbf{\Phi}$ and $\mathbf{\Psi}$.

Definition 2 (Well-conditioned density clusters)**.** *For $\lambda > 0$ and $\mathcal{C} \in \mathbb{C}_f(\lambda)$, let $\mathcal{C}$ satisfy (A1) - (A4) with respect to parameters $\sigma, \lambda_\sigma, \gamma > 0$ and $\Lambda_\sigma, D < \infty$. Letting $\kappa_1(\mathcal{C})$ and $\kappa_2(\mathcal{C})$ be given by*

$$\begin{aligned}\kappa_1(\mathcal{C}) &:= \frac{\mathbf{\Phi}(\sigma,\lambda,\lambda_\sigma,\gamma)}{\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)} \\ \kappa_2(\mathcal{C}) &:= \kappa_1(\mathcal{C})\cdot\sqrt{\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)},\end{aligned}$$

we call $\mathcal{C}$ a (κ_1,κ_2)-well-conditioned density cluster (with respect to $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ and D).

$\mathbf{\Phi}$ and $\mathbf{\Psi}$ are familiar; they are exactly the upper and lower bounds on $\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])$ and $\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])$ derived in Theorems 1 and 2, respectively.

Remark 7. For convenience and maximum generality, we define $\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)$ to correspond with the bound given by (13), and assume only (A1) - (A4). However, if we additionally have (A5), then we could sharpen $\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)$ to the tighter rate of (14), with nothing changing hereafter.

As is typical in the local clustering literature, our results will be stated with respect to specific choices or ranges of each of the user-specified parameters, which in this case may depend on the underlying (unknown) density.

In particular, for a well conditioned density cluster $\mathcal{C}$ (with respect to some $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ and D), we require

$$\begin{aligned}\alpha \in [1/10, 1/9]\cdot\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D),\, r \leq \sigma/4d \\ \pi_0 \in [2/3, 6/5]\frac{\lambda_\sigma}{\nu(\mathcal{C}_\sigma)\Lambda_\sigma^2},\, v \in \mathcal{C}_\sigma[\mathbf{X}]^g\end{aligned} \qquad (16)$$

where $\mathcal{C}_\sigma[\mathbf{X}]^g \subseteq \mathcal{C}_\sigma[\mathbf{X}]$ is some 'good' subset of $\mathcal{C}_\sigma[\mathbf{X}]$ which, as we will see, satisfies $\mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}]^g) \geq \mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}])/2$. (Intuitively one can think of $\mathcal{C}_\sigma[\mathbf{X}]^g$ as being the nodes sufficiently close to the center of $\mathcal{C}_\sigma[\mathbf{X}]$, although we provide no formal justification to this effect.)

Definition 3. *If the input parameters to Algorithm 1 satisfy 16 with respect to some $\mathcal{C}_\sigma[\mathbf{X}]$, we say the algorithm is* well-initialized.

Theorem 4. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a (κ_1,κ_2)-well conditioned cluster (with respect to some $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ and D). If*

$$\kappa_2 \leq \frac{1}{40\cdot 36}\frac{\lambda_\sigma^2}{\Lambda_\sigma^2}\frac{r^d\nu_d}{\nu(\mathcal{C}_\sigma)}. \qquad (17)$$

and Algorithm 1 is well-initialized, the output set $\widehat{C} \subset \mathbf{X}$ is a consistent estimator for $\mathcal{C}$, in the sense of Definition 1.

Remark 8. We note that larger constants than $40\cdot 36$ allow for wider range of the parameters in (16).

Approximate cluster recovery via PPR. In (Zhu et al., 2013), building on the work of (Andersen et al., 2006) and others, theory is developed which links algorithmic performance of PPR to the normalized cut and mixing time parameters. Although not the primary focus of our work, it is perhaps worth noting that these results, coupled with Theorems 1-3, translate immediately into bounds on the normalized cut and symmetric set difference of $\widehat{C}$.

We collect some of the main results of (Zhu et al., 2013) in Lemma 1.

For $G = (V,E)$ consider some $A \subseteq V$, and let $\Phi(A;G)$ and $\Psi(A;G)$ be defined as in (2) and (4), respectively.

Lemma 1 (PPR clustering)**.** *There exists a set $A^g \subset A$ with $\mathrm{vol}(A^g;G) \geq \mathrm{vol}(A;G)/2$ such that the following statement holds: Choose any $v \in A^g$, fix $\alpha = 9/10\Psi(A;G)$, and compute the page rank vector $\mathbf{p}(v,\alpha;G)$. Letting*

$$\widehat{C} = \arg\min_{\beta\in[\frac{1}{8},\frac{1}{2}]}\Phi(S_\beta;G)$$

the following guarantees hold:

$$\mathrm{vol}(\widehat{C} \setminus A) \leq \frac{24\Phi(A;G)}{\Psi(A;G)}\mathrm{vol}(A)$$

$$\mathrm{vol}(A \setminus \widehat{C}) \leq \frac{30\Phi(A;G)}{\Psi(A;G)}\mathrm{vol}(A)$$

$$\Phi(\widehat{C};G) = O\left(\frac{\Phi(A;G)}{\sqrt{\Psi(A;G)}}\right)$$

Corollary 1 immediately follows.

Corollary 1. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a (κ_1, κ_2)-well conditioned cluster (with respect to some $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ and D). Then, if Algorithm 1 is well-initialized (in the sense that the choices of input parameters satisfy (16)), the following guarantees hold for output set $\widehat{C} \subset \mathbf{X}$:*

$$\mathrm{vol}(\widehat{C} \setminus \mathcal{C}_\sigma[\mathbf{X}]), \mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}] \setminus \widehat{C}) \leq 30\kappa_1(\mathcal{C})\mathrm{vol}(\mathcal{C}_\sigma[X])$$
$$\Phi_{n,r}(\widehat{C}) = O\left(\kappa_2(\mathcal{C})\right)$$

4. Examples

Example 1 is intended to show how the machinery developed above translates in a specific, common mixture model, and the extent to which bounds are (or are not) tight. Example 2 will try to delve into some of the details of how PPR interpolates the conductance and density cut, and will show a case where a poorly conditioned density cluster is not recovered by PPR . Example 3 will emphasize finite sample cluster recovery, for a well-conditioned but non-convex mixture model

Examples 1 and 2 should be thought of as shedding light on the population performance of PPR, whereas Example 3 shows performance on a finite sample.

1. *Gaussian Mixture Model:* We will compute optimal $\mathbf{\Phi}$ and $\mathbf{\Psi}$ for given λ, and show the following
   - A graph comparing $\mathbf{\Phi}$ to $\Phi_{n,r}$ as the value of λ changes.
   - A graph comparing $\mathbf{\Psi}$ to $\Psi_{n,r}$ as the value of λ changes.
   - That for some values of λ, the conditions required for Theorem 4 hold.
2. *Thin and long parallel clusters, with ϵ-uniform noise:* We will (try to) show that the set outputted by PPR interpolates between the minimum normalized cut solution (fatter) and the density cluster (thinner). The *conductance* is

   $$\Phi^\star(G_{n,r}) := \min_{C \subset \mathbf{X}} \Phi_{n,r}(C)$$

   and the conductance cut is $C^\star \subset \mathbf{X}$ which achieves the minimum.

   We will show that
   - For sufficiently small ϵ, all three of the conductance cut, PPR cut, and density cut agree.
   - For an intermediate value of ϵ, the conductance cut and the density cut disagree. The PPR cut interpolates between the two.
   - For a sufficiently large value of ϵ, the PPR cut fails to recover the density cut, and draws closer to the conductance cut.
3. *Non-convex mixture model:* We will show that, for well-conditioned non-convex mixture model, and a finite sample size n, cluster recovery is achieved with high probability over repeated simulations.

5. Discussion

For a clustering algorithm and a given object (such as a graph or set of points), there are an almost limitless number of ways to define what the 'right' clustering is. We have considered a few such ways – density level sets, and the bicriteria of normalized cut, inverse mixing time – and shown that under the right conditions, the latter agree with the former, with resulting algorithmic consequences.

There are still many directions worth pursuing in this area. Concretely, we might wish to generalize our results to hold over a wider range of kernel functions, and hyperparameter inputs to the PPR algorithm. More broadly, we do not provide any sort of theoretical lower bound, although we give empirical evidence in Example 2 that poorly conditioned density clusters are not consistently estimated by PPR . Example 2 also hints at a way of understanding local spectral algorithms – or, at least, PPR– as interpolating between normalized and density cut. Exploring this connection is an avenue for future work.

References

Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., and Malek, A. Hit-and-run for sampling and planning in non-convex spaces. In *Artificial Intelligence and Statistics*, pp. 888–895, 2017.

Abbe, E. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.

Andersen, R., Chung, F., and Lang, K. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 475–486, 2006.

Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for the cluster tree. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 343–351. Curran Associates, Inc., 2010.

Chaudhuri, K., Graham, F. C., and Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. In *COLT*, volume 23, pp. 35.1–35.23, 2012.

Donath, W. E. and Hoffman, A. J. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, 17(5):420–425, September 1973.

Ery, A.-C., Pelletier, B., and Pudlo, P. The normalized graph cut and cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44(4):907–937, 12 2012.

Fiedler, M. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010. ISSN 0370-1573.

Hartigan, J. A. Consistency of single-linkage for high-density clusters. *Journal of the American Statistical Association*, 1981.

Haveliwala, T. H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

Hein, M., Audibert, J.-Y., and von Luxburg, U. From graphs to manifolds – weak and strong pointwise consistency of graph laplacians. In *Learning Theory*, 2005.

Jerrum, M. and Sinclair, A. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004. ISSN 0004-5411.

Koltchinskii, V. and Gine, E. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 02 2000.

Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015.

Leskovec, J., Lang, K. J., and Mahoney, M. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Mahoney, M. W., Orecchia, L., and Vishnoi, N. K. A local spectral method for graphs: with applications to improving graph partitions and exploring data graphs locally. *Journal of Machine Learning Research*, 13:2339–2365, 2012.

Maier, M., von Luxburg, U., and Hein, M. How the result of graph clustering methods depends on the construction of the graph. *CoRR*, abs/1102.2075, 2011.

McSherry, F. Spectral partitioning of random graphs. In *FOCS*, pp. 529–537, 2001.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 08 2011.

Schiebinger, G., Wainwright, M. J., and Yu, B. The geometry of kernelized spectral clustering. *Ann. Statist.*, 43(2): 819–846, 04 2015.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.*, 37(6B):3960–3984, 12 2009.

Singh, A., Scott, C., and Nowak, R. Adaptive hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B): 2760–2782, 10 2009.

Spielman, D. A. and Teng, S.-H. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Spielman, D. A. and Teng, S.-H. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

Spielman, D. A. and Teng, S.-H. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

Tolliver, D. and Miller, G. L. Graph partitioning by spectral rounding: Applications in image segmentation and clustering. In *Computer Vision and Pattern Recognition,CVPR*, volume 1, pp. 1053–1060, 2006.

Trillos, N. G., Slepčev, D., Von Brecht, J., Laurent, T., and Bresson, X. Consistency of cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 17(1):6268–6313, January 2016.

Vempala, S. and Wang, G. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841 – 860, 2004.

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

von Luxburg, U., Belkin, M., and Bousquet, O. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 04 2008.

Yang, J. and Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, Jan 2015.

Zhu, Z. A., Lattanzi, S., and Mirrokni, V. S. A local algorithm for finding well-connected clusters. In *ICML (3)*, pp. 396–404, 2013.