Local Spectral Clustering of Density Upper Level Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Spectral clustering methods are a family of popular nonparametric 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. In contrast to existing works, we analyze PPR in a traditional statistical learning setup, where we obtain samples from an unknown distribution, and aim to identify connected regions of high-density (density clusters). We prove that PPR, run on a neighborhood graph, extracts sufficiently salient density clusters, and provide empirical support of our theory.

1 Introduction

Let $X = \{x_1, \ldots, x_n\}$ be a sample drawn i.i.d. from a distribution $\mathbb{P}$ on $\mathbb{R}^d$, with density f, and consider the problem of clustering: splitting the data into groups which satisfy some notion of within-group similarity and between-group difference. We focus on spectral clustering methods, a family of powerful nonparametric clustering algorithms. Roughly speaking, a spectral technique first constructs a geometric graph G, where vertices are associated with samples, and edges correspond to proximities between samples. It then learns a feature embedding based on the Laplacian of G, and applies a simple clustering technique (such as k-means clustering) in the embedded feature space.

To be more precise, let $G = (V, E, w)$ denote a weighted, undirected graph constructed from the samples X, where $V = \{1, \ldots, n\}$, and $w_{uv} = K(x_u, x_v) \geq 0$ for $u, v \in V$, and a particular kernel function K. Here $(u, v) \in E$ if and only if $w_{uv} > 0$. We denote by $\mathbf{A} \in \mathbb{R}^{n \times n}$ the weighted adjacency matrix, which has entries $A_{uv} = w_{uv}$, and by $\mathbf{D}$ the degree matrix, with $\mathbf{D}_{uu} = \sum_{v \in V} \mathbf{A}_{uv}$. We also denote by $\mathbf{W}, \mathbf{L}$ the (lazy) random walk transition probability matrix and normalized[1] Laplacian matrix, respectively, which are defined as

$$\mathbf{W} = \frac{\mathbf{I} + \mathbf{D}^{-1}\mathbf{A}}{2}, \quad \mathbf{L} = \mathbf{I} - \mathbf{W},$$

where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. Classical global spectral methods take a eigendecomposition $L = U \Sigma U^T$, use some number of eigenvectors (columns in U) as a feature representation for the samples, and then run (say) k-means in this new feature space.

When applied to geometric graphs constructed from a large number of samples, global spectral clustering methods can be computationally cumbersome and insensitive to the local geometry of the underlying distribution [Leskovec et al., 2010, Mahoney et al., 2012]. This has led to recent increased interest in local spectral algorithms, which leverage locally-biased spectra computed using random walks around a user-specified seed node. A popular local clustering algorithm is Personalized PageRank (PPR), first introduced by [Haveliwala, 2003], and further developed by [Spielman and Teng, 2011, 2014, Andersen et al., 2006, Mahoney et al., 2012, Zhu et al., 2013], among others.

[1] Other popular choices here include the unnormalized Laplacian, and symmetric normalized Laplacian.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Local spectral clustering techinques have been practically very successful [Leskovec et al., 2010, Andersen et al., 2012, Gleich and Seshadhri, 2012, Mahoney et al., 2012, Wu et al., 2012], which has led many authors to develop supporting theory [Spielman and Teng, 2013, Andersen and Peres, 2009, Gharan and Trevisan, 2012, Zhu et al., 2013] that gives worst-case guarantees on traditional graph-theoretic notions of cluster quality (like conductance). In this paper, we adopt a more traditional statistical viewpoint, and examine what the output of a local clustering algorithm on X reveals about the unknown density f. In particular, we examine the ability of the PPR algorithm to recover *density clusters* of f, which are defined as the connected components of the upper level set $\{x \in \mathbb{R}^d : f(x) \geq \lambda\}$ for some threshold $\lambda > 0$ (a central object of central interest in the classical statistical literature on clustering, dating back to Hartigan [1981]).

1.1 PPR on a Neighborhood Graph

We now describe the clustering algorithm that will be our focus for the rest of the paper. We start with the geometric graph that we form based on the samples X: for a radius $r > 0$, we consider the *r-neighborhood graph* of X, denoted $G_{n,r} = (V, E)$, an unweighted graph with vertices $V = X$, and an edge $(x_i, x_j) \in E$ if and only if $\|x_i - x_j\| \leq r$, where $\|\cdot\|$ denotes Euclidean norm. Note that this is a special case of the general construction introduced above, with $K(u, v) = 1(\|x_u - x_v\| \leq r)$.

Next, we define the PPR vector $p = p(v, \alpha; G_{n,r})$, with respect to a seed node $v \in V$ and a teleportation parameter $\alpha \in [0, 1]$, to be the solution of the following linear system:

$$p = \alpha \mathbf{e}_v + (1 - \alpha) p \mathbf{W}, \tag{1}$$

where $\mathbf{W}$ is the random walk matrix of the underlying graph $G_{n,r}$ and e_v denotes indicator vector for node v (with a 1 in the vth position and 0 elsewhere). In practice, we can approximately solve the above linear system via a simple, efficient random walk, with appropriate restarts to v.

For a level $\beta > 0$ and a target volume $\mathrm{vol}_0 > 0$, we define a *β-sweep cut* of $p = (p_u)_{u \in V}$ as

$$S_\beta = \{u \in V : \frac{p_u}{\mathbf{D}_{uu}} > \frac{\beta}{\mathrm{vol}_0}\}. \tag{2}$$

Having computed sweep cuts S_β over a range $\beta \in (\frac{1}{40}, \frac{1}{11})$,[2], we then output a cluster estimate $\widehat{C} = S_{\beta^*}$ to have minimum normalized cut $\Phi(S_{\beta^*}; G_{n,r})$, where for $S \cup S^c = G_{n,r}$, $\mathrm{cut}(S; G_{n,r}) := |\{(u, v) \in E : u \in S, v \in S^c\}|$, $\mathrm{vol}(S; G_{n,r}) := \sum_{u \in S} \mathbf{D}_{uu}$, and

$$\Phi(S; G_{n,r}) := \frac{\mathrm{cut}(S; G_{n,r})}{\min\{\mathrm{vol}(S; G_{n,r}), \mathrm{vol}(S^c; G_{n,r})\}}. \tag{3}$$

For concreteness, we summarize this procedure in Algorithm 1.

Algorithm 1 PPR on a Neighborhood Graph

Input: data $X = \{x_1, \ldots, x_n\}$, radius $r > 0$, teleportation parameter $\alpha \in [0, 1]$, seed $v \in X$, target stationary volume $\mathrm{vol}_0 > 0$.

Output: cluster $\widehat{C} \subseteq V$.

1: Form the neighborhood graph $G_{n,r}$.
2: Compute the PPR vector $p(v, \alpha; G_{n,r})$ as in (1).
3: For $\beta \in (\frac{1}{40}, \frac{1}{11})$ compute sweep cuts S_β as in (2).
4: Return $\widehat{C} = S_{\beta^*}$, where

$$\beta^* = \operatorname*{arg\,min}_{\beta \in (\frac{1}{40}, \frac{1}{11})} \Phi(S_\beta; G_{n,r}).$$

[2]The choice of a specific range such as $(\frac{1}{40}, \frac{1}{11})$ is standard in the analysis of PPR algorithms, see, e.g., [Zhu et al., 2013].

1.2 Summary of Results

Let $\mathbb{C}_f(\lambda)$ denote the connected components of the density upper level set $\{x \in \mathbb{R}^d : f(x) > \lambda\}$. For a given density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we call $\mathcal{C}[X] = \mathcal{C} \cap X$ the *empirical density cluster*. Below we give two notions of performance of a density cluster estimate.

Definition 1 (Misclassification error)**.** *For an estimator $\widehat{C} \subseteq X$ and set $\mathcal{S} \subseteq \mathbb{R}^d$, the* misclassification error *of $\mathcal{S}$ by $\widehat{C}$ is*

$$|\widehat{C} \setminus (\mathcal{S} \cap X)| + |(\mathcal{S} \cap X) \setminus \widehat{C}|. \tag{4}$$

Definition 2 (Consistent density cluster estimation)**.** *For an estimator $\widehat{C} \subseteq X$ and cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we say $\widehat{C}$ is a consistent estimator of $\mathcal{C}$ if for all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ with $\mathcal{C} \neq \mathcal{C}'$ the following holds as $n \to \infty$:*

$$\mathcal{C}[X] \subseteq \widehat{C} \quad \textit{and} \quad \widehat{C} \cap \mathcal{C}'[X] = \emptyset, \tag{5}$$

with probability tending to 1.

A summary of our main results (and outline for the rest of this paper) is as follows.

1. In Section 2, we introduce a set of natural geometric conditions. We formalize a measure of difficulty based on these geometric conditions, and show that when properly initialized, the misclassification error of Algorithm 1 is upper bounded by this difficulty measure.
2. We further show that if the density cluster $\mathcal{C}$ is particularly well-conditioned, Algorithm 1 will perform consistent density cluster estimation in the sense of (5).
3. Corollary 1 establishes that these statements hold also with respect to an approximate form of PPR, which can be efficiently computed.
4. In Section 3, we detail some of the main technical machinery required to prove our main results, highlighting the part various geometric quantities play in the ultimate difficulty of the clustering problem.
5. In Section 4, we empirically demonstrate the tightness of the bounds in Theorems 3 and 4, and provide examples showing how violations of the geometric conditions we require manifestly impact density cluster recovery by PPR.

On the topic of conditions, it is worth mentioning that, as density clusters are inherently local, focusing on the PPR algorithm actually eases our analysis and allows us to require fewer global regularity conditions relative to those needed for more classical global spectral algorithms.

1.3 Related Work

In addition to the background given above, a few related lines of work are worth highlighting. Building on earlier work of [Koltchinskii and Gine, 2000], [von Luxburg et al., 2008, Hein et al., 2005] studied the limiting behaviour of spectral clustering algorithms. These authors show that when samples are obtained from a distribution, and we appropriately construct a geometric graph, 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.

More similar to our results are the works [Vempala and Wang, 2004, Shi et al., 2009, Schiebinger et al., 2015], who study the consistency of spectral algorithms in recovering the latent labels in certain parametric and nonparametric 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 the focus in our work.

2 Estimation of Well-Conditioned Density Clusters.

2.1 Geometric Conditions on Density Clusters

As mentioned previously, successful recovery of a density cluster by PPR requires the density cluster to be geometrically well-conditioned. At a minimum, we wish to avoid dumbbell-like sets $\mathcal{C}$ which

contain an arbitrarily thin bridge, and as in Chaudhuri and Dasgupta [2010] we therefore introduce a buffer zone around $\mathcal{C}$. Letting $B(x,r)$ be the closed ball of radius $r > 0$ centered at $x \in \mathbb{R}^d$, for a given cluster $\mathcal{C} \subseteq \mathbb{R}^d$ and $\sigma > 0$, we refer to $\mathcal{C}_\sigma := \left\{y \in \mathbb{R}^d : \inf_{x \in \mathcal{C}} \|y - x\| \leq \sigma\right\}$ as the σ-expansion of $\mathcal{C}$, and state our conditions with respect to $\mathcal{C}_\sigma$.

More generally, over the neighborhood graph $G_{n,r}$ we would like the empirical cluster $\mathcal{C}_\sigma[X]$ to be well connected everywhere in its interior, and poorly connected to the rest of X. This intuition motivates our required conditions, stated with respect to a density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$ for some threshold $\lambda > 0$, and an expansion parameter $\sigma > 0$.

(A1) *Bounded density within cluster:* There are $0 < \lambda_\sigma < \Lambda_\sigma < \infty$ such that

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

(A2) *Cluster separation:* For all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ with $\mathcal{C}' \neq \mathcal{C}$,

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

where $\operatorname{dist}(\mathcal{C}, \mathcal{C}') = \inf_{x \in \mathcal{C}} \operatorname{dist}(x, \mathcal{C}')$.

(A3) *Low noise density:* There exists $\gamma, c_0 > 0$ such that for all $x \in \mathbb{R}^d$ with $0 < \operatorname{dist}(x, \mathcal{C}_\sigma) \leq \sigma$,

$$\inf_{x' \in \mathcal{C}_\sigma} f(x') - f(x) \geq c_0 \operatorname{dist}(x, \mathcal{C}_\sigma)^\gamma,$$

where $\operatorname{dist}(x, \mathcal{C}) = \inf_{x_0 \in \mathcal{C}} \|x - x_0\|$.

(A4) *Lipschitz embedding:* $\mathcal{C}_\sigma$ is the image of a convex set under a biLipschitz, measure preserving mapping. Formally, there exists $\mathcal{K} \subseteq \mathbb{R}^d$ convex, and $g : \mathbb{R}^d \to \mathbb{R}^d$ such that $\det(\nabla g(x)) = 1$ for all $x \in \mathcal{C}_\sigma$, and for some $L \geq 1$,

$$\frac{1}{L}\|x - y\| \leq \|g(x) - g(y)\| \leq L\|x - y\| \text{ for all } x, y \in \mathcal{C}_\sigma$$

such that $\mathcal{C}_\sigma$ is the image of $\mathcal{K}$ by g, $\mathcal{C}_\sigma = g(\mathcal{K})$. Furthermore, there exists $D < \infty$ such that for all $x, x' \in \mathcal{K}$

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

(A5) *Bounded volume:* Let the neighborhood graph radius $0 < r \leq \sigma/2d$ be such that

$$\frac{\int_{\mathcal{C}_\sigma} \mathbb{P}(B(x,r)) f(x) dx}{\int_{\mathbb{R}^d} \mathbb{P}(B(x,r)) f(x) dx} \leq \frac{1}{2}. \tag{6}$$

The cluster separation (A2) and low noise density (A3) conditions guarantee poor connectivity between $\mathcal{C}_\sigma[X]$ and $X \setminus \mathcal{C}_\sigma[X]$, whereas (A1) and (A4) ensure high connectivity within $\mathcal{C}_\sigma[X]$. It may not be immediately obvious how (A4) contributes to geometric conditioning. For now, we observe merely that random walks will mix slowly over sets with large diameter, and make some more detailed commentary in Section 3. Finally, (A5) is a relatively harmless technical condition, merely excluding the case where $\mathcal{C}_\sigma$ contains over half the total mass.

2.2 Well-Conditioned Density Clusters

We turn to formally defining a condition number, $\kappa(\mathcal{C})$, reflects the difficulty of the local spectral clustering task. The smaller $\kappa(\mathcal{C})$ is, the more success PPR will have in recovering $\mathcal{C}$. Let $\theta := (r, \sigma, \lambda, \lambda_\sigma, \Lambda_\sigma, \gamma, D, L)$ contain those geometric parameters detailed in 2.1.

Definition 3 (Well-conditioned density clusters)**.** *For $\lambda > 0$ and $\mathcal{C} \in \mathbb{C}_f(\lambda)$, let $\mathcal{C}$ satisfy (A1) - (A5) for some θ, and additionally let $\mathcal{C}_\sigma$ satisfy (6) . Then, setting*

$$\begin{aligned}
\mathbf{\Phi}(\theta) &:= c_1 r \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} \\
\mathbf{\Psi}(\theta) &:= \left(c_2 \frac{\Lambda_\sigma^4 d^3 D^2 L^2}{\lambda_\sigma^4 r^2} \log^2\left(\frac{1}{r}\right) + c_3 \log\left(\frac{\Lambda_\sigma}{\lambda_\sigma}\right) \right)^{-1}
\end{aligned} \tag{7}$$

and

$$\kappa(\mathcal{C}) := \frac{\mathbf{\Phi}(\theta)}{\mathbf{\Psi}(\theta)} \tag{8}$$

we call $\mathcal{C}$ a κ-well-conditioned density cluster.

At first glance (7) may appear mysterious, but as will be shown in Section 3, these are merely upper bounds on the normalized cut and inverse mixing time of (the σ-expansion of) a given empirical density cluster $\mathcal{C}_\sigma[X]$ in $G_{n,r}$. In Zhu et al. [2013], building on the work of Andersen et al. [2006] and others, it is shown that the ratio of normalized cut to inverse mixing time is a fundamental quantity governing the performance of PPR over a general graph. $\kappa(\mathcal{C})$ upper bounds this ratio for an empirical density cluster over the neighborhood graph $G_{n,r}$, and is therefore a natural criterion to measure difficulty of the clustering task.

Well-initialized algorithms. As is typical in the local clustering literature, our algorithmic results will be stated with respect to specific choices or ranges of each of the user-specified parameters.

In particular, for a well-conditioned density cluster $\mathcal{C}$ (with respect to some θ), we require

$$\begin{aligned} r \leq \frac{\sigma}{2d}, \alpha \in [1/10, 1/9] \cdot \mathbf{\Psi}(\theta), \\ v \in \mathcal{C}_\sigma[X]^g, \mathrm{vol}_0 \in [3/4, 5/4] \cdot n(n-1) \int_{\mathcal{C}_\sigma} \mathbb{P}(B(x,r)) f(x) dx \end{aligned} \tag{9}$$

$\mathcal{C}_\sigma[X]^g \subseteq \mathcal{C}_\sigma[X]$ will be some large subset of $\mathcal{C}_\sigma[X]$, in particular $\mathrm{vol}(\mathcal{C}_\sigma[X]^g; G_{n,r}) \geq \mathrm{vol}(\mathcal{C}_\sigma[X]; G_{n,r})/2$.

Definition 4. *If the input parameters to Algorithm 1 satisfy* (9) *for some well-conditioned density cluster $\mathcal{C}$, we say the algorithm is* well-initialized.

In practice it is clearly not feasible to set hyperparameters based on the underlying (unknown) density f. Typically, one tunes PPR over a range of hyperparameters and optimizes for some criterion such as normalized cut; it is unclear how this scheme would affect the performance of PPR in the density clustering context.

Density cluster estimation by PPR. Theorem 1 of Zhu et al. [2013], combined with the results of Section 3, immediately implies a bound on the volume of $\widehat{C} \setminus \mathcal{C}_\sigma[X]$ (and likewise $\mathcal{C}_\sigma[X] \setminus \widehat{C}$),

$$\mathrm{vol}_{n,r}(\widehat{C} \setminus \mathcal{C}_\sigma[X]), \mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X] \setminus \widehat{C}) \lesssim \kappa(\mathcal{C}) \mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X]). \tag{10}$$

To translate (10) into meaningful bounds on misclassification error, we wish to preclude vertices $x \in X$ from having arbitrarily small degree. To do so, we make some regularity assumptions on $\mathcal{X} := \mathrm{supp}(f)$.

(A5) *Valid region:* There exists some number $\lambda_{\min} > 0$ such that $\lambda_{\min} < f(x)$ for all $x \in \mathcal{X}$. Additionally, there exists some $c > 0$ such that for each $x \in \partial\mathcal{X}$, $\nu(B(x,r) \cap \mathcal{X}) \geq c\nu(B(x,r))$.

Note that the latter condition in $(A5)$ will be satisfied if, for instance, $\mathcal{X}$ is a σ-expanded set.

Theorem 1. *Fix $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a κ-well conditioned density cluster (with respect to some θ), and additionally assume f satisfies (A5). Then, with probability tending to one as $n \to \infty$,*

$$\frac{|\mathcal{C}_\sigma[X] \setminus \widehat{C}|}{\left|\mathcal{C}_\sigma[X]\right|} \leq c_5 \kappa(\mathcal{C}) \frac{\Lambda_\sigma}{\lambda_\sigma}, \quad \textit{and} \quad \frac{|\widehat{C} \setminus \mathcal{C}_\sigma[X]|}{\left|\mathcal{C}_\sigma[X]\right|} \leq c_6 \kappa(\mathcal{C}) \frac{\Lambda_\sigma}{\lambda_{\min}}. \tag{11}$$

for universal constants $c_4, c_5 > 0$.

The proof of Theorem 1, along with all other proofs in this paper, can be found in the supplementary material. We observe that the misclassification error is proportional to the difficulty of the clustering problem, as measured by the condition number.

Neither (10) nor Theorem 1 imply consistent density cluster estimation in the sense of (5). This notion of consistency requires a uniform bound over p for all $u \in \mathcal{C}, u' \in \mathcal{C}'$

$$\frac{p_{u'}}{\mathbf{D}_{uu}} \leq \frac{1}{40\mathrm{vol}_0} < \frac{1}{11\mathrm{vol}_0} \leq \frac{p_u}{\mathbf{D}_{uu}}. \tag{12}$$

so that any sweep cut S_β for $\beta\mathrm{vol}_0 \in [1/40, 1/11]$ (i.e. any sweep cut considered by Algorithm 1) will fulfill both conditions laid out in (5). In Theorem 2, we show that a sufficiently small upper bound on $\kappa(\mathcal{C})$ ensures such a gap exists with probability one as $n \to \infty$, and therefore guarantees $\widehat{C}$ will be a consistent estimator. As was the case before, we wish to preclude arbitrarily low degree vertices, this time for points $x \in \mathcal{C}'[X]$.

(A6) *$\mathcal{C}'$- bounded density :* For each $\mathcal{C}' \in \mathbb{C}_f(\lambda), \mathcal{C}' \neq \mathcal{C}$ and for all $x \in \mathcal{C}' + \sigma B$, $\lambda_\sigma \leq f(x)$ where σ, λ_σ are as in (A1).

Theorem 2. *Fix $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a κ-well conditioned cluster (with respect to some θ), and additionally assume (A6) holds. If Algorithm 1 is well-initialized, there exists universal constant $c_7 > 0$ such that if*

$$\kappa(\mathcal{C}) \leq c_7 \frac{\lambda_\sigma^2 r^d \nu_d}{\Lambda_\sigma \mathbb{P}(\mathcal{C}_\sigma)}, \tag{13}$$

then the output set $\widehat{C} \subseteq X$ is a consistent estimator for $\mathcal{C}$, in the sense of Definition 2.

A few remarks are in order.

Remark 1. We note that the restriction on $\kappa(\mathcal{C})$ imposed by (13) results in a misclassification rate on the order of r^d. (See Theorem 1). In plain terms, we are able to recover a density cluster $\mathcal{C}$ in the sense of (5) only when we can guarantee a very small fraction of points are misclassified. This strong condition is the price we pay in order to obtain the uniform bound of 12.

Remark 2. While taking the radius of the neighborhood graph $r \to 0$ as $n \to \infty$—and thereby ensuring $G_{n,r}$ is sparse—is computationally attractive, the presence of a factor of $\frac{\log^2(1/r)}{r}$ in $\kappa(\mathcal{C})$ unfortunately prevents us from making claims about the behavior of PPR in this regime. Although the restriction to a kernel function fixed in n is standard for theoretical analysis of spectral clustering Schiebinger et al. [2015], von Luxburg et al. [2008], it is an interesting question whether PPR exhibits some degeneracy over r-neighborhood graphs as $r \to 0$, or if this is merely looseness in our upper bounds.

Cluster estimation with the approximate PPR **vector.** As mentioned previously, in practice exactly solving (1) may be too computationally expensive. To address this limitation, Andersen et al. [2006] introduced the *ϵ-approximate* PPR *vector* (aPPR), which we will denote $p^{(\epsilon)}$. We refer the curious reader to Andersen et al. [2006] for a formal algorithmic definition of the aPPR vector, and limit ourselves to highlighting a few salient points. Namely, the aPPR vector can be computed in $\mathcal{O}\left(\frac{1}{\epsilon\alpha}\right)$ time, while satisfying the following uniform error bound:

$$\text{for all } x \in X, \quad p(x) - \epsilon \deg_{n,r}(x) \leq p^{(\epsilon)}(x) \leq p(x) \tag{14}$$

Application of (14) within the proofs of Theorems 1 and 2 leads to analogous results which hold with respect to $p^{(\epsilon)}$.

Corollary 1. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a κ-well-conditioned cluster (with respect to some θ). Choose input parameters $\alpha, r, \mathrm{vol}_0, v$ to be well-initialized in the sense of* (9)*, set $\epsilon = \frac{1}{20\mathrm{vol}_0}$, and modify Algorithm 1 to compute the a*PPR *vector $p^{(\epsilon)}$ rather than the exact* PPR *vector p, with resulting output $\widehat{C}$.*

1. *Assume (A5) holds. Then* (11) *is still a valid upper bound for the misclassification error of $\widehat{C}$.*
2. *Assume (A6) holds. If*

$$\kappa(\mathcal{C}) \leq c_7 \frac{\lambda_\sigma^2}{\Lambda_\sigma^2} \frac{r^d \nu_d}{\nu(\mathcal{C}_\sigma)}$$

then $\widehat{C} \subseteq X$ is a consistent estimator for $\mathcal{C}$, in the sense of Definition 2.

3 Analysis

Given an arbitrary graph $G = (V, E)$ and candidate cluster $S \subseteq G$, Zhu et al. [2013] bound the volume of $\widehat{C} \setminus S$ and $S \setminus \widehat{C}$ in terms of the normalized cut and inverse mixing time of S. The key to deriving the algorithmic results of the previous section is therefore to show that the geometric conditions (A1) - (A4) translate to meaningful bounds on the normalized cut and inverse mixing time of $\mathcal{C}_\sigma[X]$ in $G_{n,r}$. Doing so constitutes the bulk of our technical effort.

3.1 Upper Bound on Normalized Cut

We start with an upper bound on the normalized cut (3) of $\mathcal{C}_\sigma[X]$. (In Theorem 3, the upper bound on the density in Assumption (A1) will not actually be needed, so we omit the parameter $\Lambda_\sigma > 0$ from the theorem statement.) For simplicity, we write $\Phi_{n,r}(\mathcal{C}_\sigma[X]) := \Phi(\mathcal{C}_\sigma[X]; G_{n,r})$.

Theorem 3. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1)-(A3), and (A5) for some $r, \sigma, \lambda_\sigma, c_0, \gamma > 0$. Then for any $0 < \delta < 1$, $\epsilon > 0$, if*

$$n \geq \frac{(2+\epsilon)^2 \log(3/\delta)}{\epsilon^2} \left(\frac{25}{6\lambda_\sigma^2 \nu(\mathcal{C}_\sigma) \nu_d r^d} \right)^2, \tag{15}$$

then

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[X])}{r} \leq c_1 \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon, \tag{16}$$

with probability at least $1 - \delta$ (where $c_1 > 0$ is a universal constant).

Remark 3. Observe that the diameter D is absent from Theorem 3, in contrast to the difficulty function $\kappa(\mathcal{C})$, which worsens (increases) as D increases. This phenomenon reflects established wisdom regarding spectral partitioning algorithms more generally Guattery and Miller [1995], Hein and Bühler [2010], albeit newly applied to the density clustering setting. It suggests that PPR may fail to recover $\mathcal{C}_\sigma[X]$ even when $\mathcal{C}$ is sufficiently well-conditioned to ensure $\mathcal{C}_\sigma[X]$ has a small normalized cut in $G_{n,r}$, if the diameter D is large. This intuition will be supported by simulations in Section 4.

3.2 Lower Bound on Inverse Mixing Time

For $S \subseteq V$, denote by $G[S] = (S, E_S, w_S)$ the subgraph induced by S (where the edges are $E_S = E \cap (S \times S)$), let $\mathbf{W}_S$ be the (lazy) random walk matrix over $G[S]$, and write

$$q_v^{(t)}(u) = e_v \mathbf{W}_S^t e_u$$

for the t-step transition probability of a random walk over $G[S]$ originating at v.[3] Also write $\pi = (\pi(u))_{u \in S}$ for the stationary distribution of this random walk. (Given the definition of $\mathbf{W}_S$, it is well-known that a unique stationary distribution exists and is given by $\pi(u) = \deg(u; G[S])/\mathrm{vol}(S; G[S])$.)

Then, the *relative pointwise mixing time* of $G[S]$ is

$$\tau_\infty(G[S]) = \min \left\{ t : \frac{\pi(u) - q_v^{(t)}(u)}{\pi(u)} \leq \frac{1}{4}, \text{ for } u, v \in V \right\}. \tag{17}$$

We lower bound the inverse mixing time $\Psi_{n,r}(\mathcal{C}_\sigma[X]) = 1/\tau_\infty(\mathcal{C}_\sigma[X])$ of $\mathcal{C}_\sigma[X]$, or equivalently we upper bound the mixing time.

Theorem 4. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1) and (A4) for some $\sigma, \lambda_\sigma, \Lambda_\sigma, D, K > 0$. Then, for any $0 < r < \sigma/2\sqrt{d}$, with probability one*

$$\limsup_{n \to \infty} \tau_\infty(\mathcal{C}_\sigma[X]) \leq c_2 \frac{\Lambda_\sigma^4 d^3 D^2 L^2}{\lambda_\sigma^4 r^2} \log^2 \left(\frac{1}{r} \right) + c_3 \log \left(\frac{\Lambda_\sigma}{\lambda_\sigma} \right) \tag{18}$$

for $c_2, c_3 > 0$ universal constants.

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

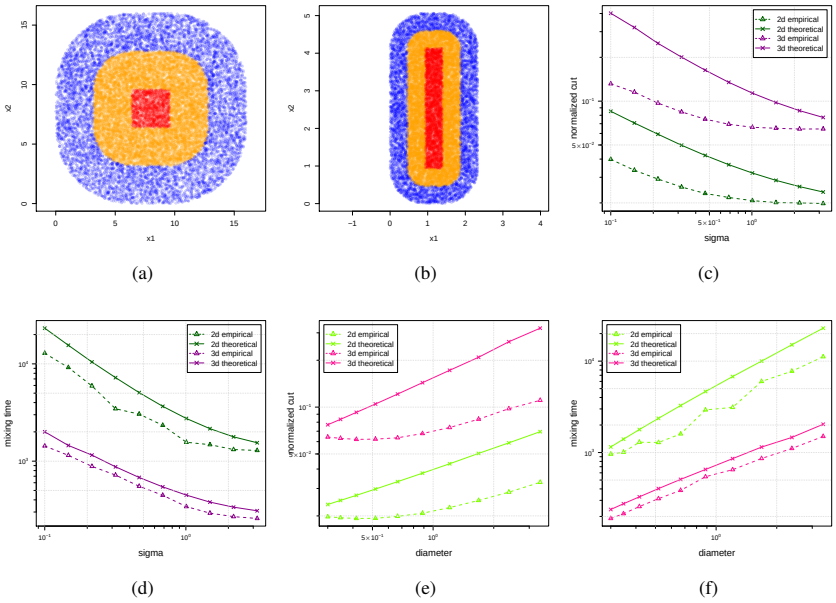

Figure 1: Samples, empirical results, and theoretical bounds for mixing time and normalized cut as diameter and thickness are varied. In (a) and (b), points in $\mathcal{C}$ are colored in red; points in $\mathcal{C}_\sigma \setminus \mathcal{C}$ are colored in yellow; and remaining points in blue.

So far as we are aware, Theorem 4 is a novel bound on the mixing time of random walks over neighborhood graphs.

Remark 4. The embedding assumption (A4) and Lipschitz parameter L play an important role in proving the upper bound of Theorem 4. There is some interdependence between L and other geometric parameters σ and D, which might lead one to hope that (A4) is non-essential. However, it is not possible to eliminate this condition without incurring an additional factor of at least $(D/\sigma)^d$ in (18), achieved, for instance, when $\mathcal{C}_\sigma$ is a dumbbell-like set consisting of two balls of diameter D linked by a cylinder of radius σ. [Abbasi-Yadkori et al., 2017, Abbasi-Yadkori, 2016] develop theory regarding biLipschitz deformations of convex sets, wherein it is observed that star-shaped sets as well as half-moon shapes of the type we consider in Section 4 both satisfy (A4) for reasonably small values of L.

4 Experiments

We provide numerical experiments to investigate the tightness of our bounds on the cluster quality criteria normalized cut and mixing time, and examine the performance of PPR on the 'two moons' dataset. For space reasons, we defer details of the experimental settings to the supplement.

Validating Theoretical Bounds. As we do not provide any theoretical lower bounds, we investigate the tightness of Theorems 3 and 4 via simulation. Figure 1 shows these theoretical bounds compared to the empirical quantities (3) and (17), as we vary the diameter D and thickness σ of the cluster $\mathcal{C}$.

Panels (d) and (f) show our theoretical bounds on mixing time tracking closely with empirical mixing time, in both 2 and 3 dimensions.[4] This provides empirical evidence that the upper bound on mixing time given by Theorem 4 has the right dependency on both expansion parameter σ and diameter D. The story in panels (c) and (e) is less obvious. We note that while, broadly speaking, the trends do not appear to match, this gap between theory and empirical results seems largest when

[4]Note that we have rescaled all values of theoretical upper bounds by a constant, in order to mask the effect of large universal constants in these bounds. Therefore only comparison of slopes, rather than intercepts, is meaningful.

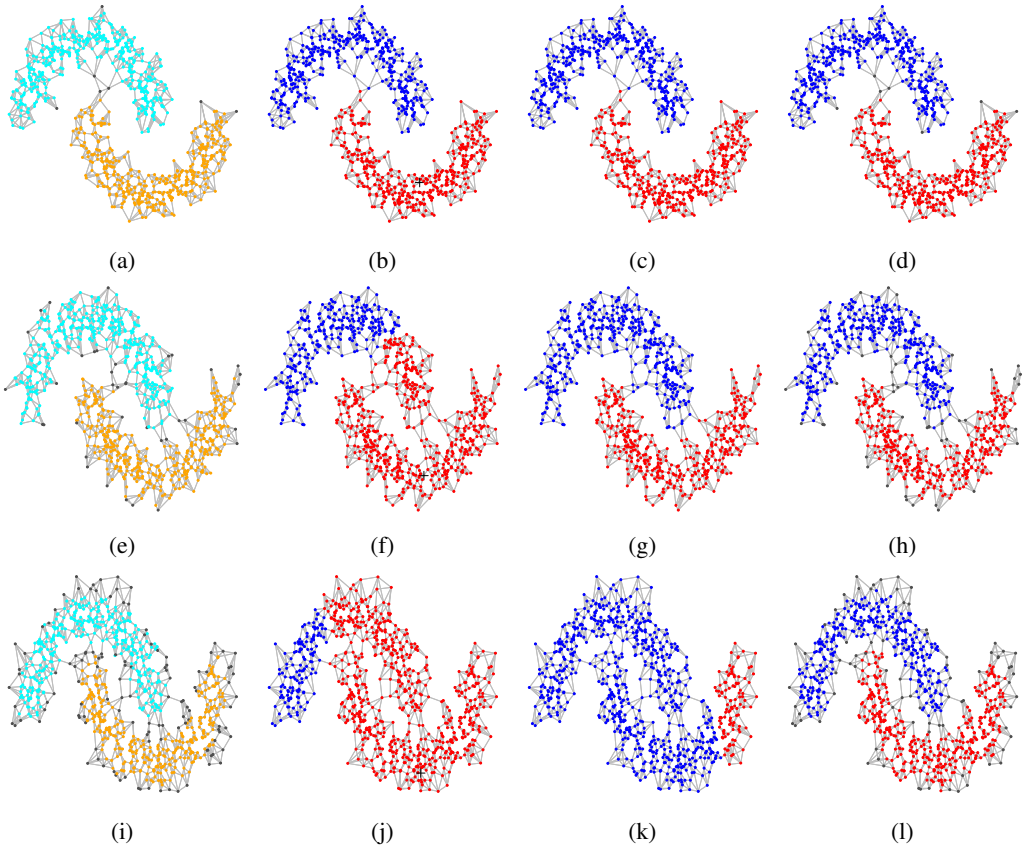

Figure 2: True density (column 1), PPR (column 2), normalized cut (column 3) and estimated density (column 4) clusters for 3 different simulated data sets. Seed node for PPR denoted by a black cross.

$\sigma \approx D$. As the ratio D/σ grows, we see the slopes of the empirical curves becoming more similar to those predicted by theory.

PPR **, normalized cut, and density clustering comparison.** To drive home the main implications of Theorems 1 and 2, in Figure 2 we show the behavior of PPR, normalized cut, and the density clustering algorithm of [Chaudhuri and Dasgupta, 2010] on (a variant of) the famous 'two moons' dataset, considered a prototypical success story for spectral clustering algorithms. The first column shows empirical density clusters C_n and C'_n for a particular threshold λ of the density function; the second column shows the cluster recovered by PPR; the third column shows the global minimum normalized cut, computed according to the algorithm of Szlam and Bresson [2010]; and the last column shows a cut of the density cluster tree estimator of Chaudhuri and Dasgupta [2010].

Figure 2 show the degrading ability of PPR to recover density clusters as the two moons become less salient. Of particular interest is the fact that PPR fails to recover one of the moons even when normalized cut still succeeds in doing so, and that a density clustering algorithm recovers a moon even when both PPR and normalized cut fail.

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.

We do not provide a theoretical lower bound showing that our geometric conditions are required for successful recovery on an upper level set. Although we investigate the matter empirically, this is a direction for future work.

References

Yasin Abbasi-Yadkori. Fast mixing random walks and regularity of incompressible vector fields. *arXiv preprint arXiv:1611.09252*, 2016.

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, and Alan Malek. Hit-and-Run for Sampling and Planning in Non-Convex Spaces. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 888–895, 2017.

Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 235–244, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-506-2. doi: 10.1145/1536414.1536449.

Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, 2006.

Reid Andersen, David F Gleich, and Vahab Mirrokni. Overlapping clusters for distributed computation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 273–282. ACM, 2012.

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

Shayan Oveis Gharan and Luca Trevisan. Approximating the expansion profile and almost optimal local graph clustering. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 187–196. IEEE, 2012.

David F Gleich and C Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 597–605. ACM, 2012.

Stephen Guattery and Gary L Miller. On the performance of spectral graph partitioning methods. In *SODA*, volume 95, pages 233–242, 1995.

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

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

Matthias Hein and Thomas Bühler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. In *Advances in Neural Information Processing Systems*, pages 847–855, 2010.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. From graphs to manifolds – weak and strong pointwise consistency of graph laplacians. In *Learning Theory*, 2005.

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

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

Michael W. Mahoney, Lorenzo Orecchia, and Nisheeth K. Vishnoi. 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.

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

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

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

Daniel A Spielman and Shang-Hua Teng. 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.

Daniel A Spielman and Shang-Hua Teng. 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.

Arthur Szlam and Xavier Bresson. Total variation, cheeger cuts. In *ICML*, pages 1039–1046, 2010.

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

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

Xiao-Ming Wu, Zhenguo Li, Anthony M. So, John Wright, and Shih fu Chang. Learning with partially absorbing random walks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3077–3085. Curran Associates, Inc., 2012.

Zeyuan Allen Zhu, Silvio Lattanzi, and Vahab S Mirrokni. A local algorithm for finding well-connected clusters. In *ICML (3)*, pages 396–404, 2013.