Local Spectral Clustering of Density Upper Level Sets

Anonymous Authors[1]

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. Several authors have shown that local methods, such as personalized PageRank (PPR), have worst-case guarantees for certain graph-based measures of cluster quality. 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 introduce two natural criteria for cluster quality, and derive bounds for these criteria when evaluated on empirical analogues of density clusters. Moreover, we prove that PPR, run on a neighborhood graph, extracts sufficiently salient density clusters.

1. Introduction

Let $\mathbf{X} = \{x_1, \ldots, x_n\}$ 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 $\mathbf{X}$, 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 random walk transition probability matrix and normalized[1] Laplacian matrix, respectively, which are defined as

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

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 & Teng (2011; 2014); Andersen et al. (2006); Mahoney et al. (2012); Zhu et al. (2013), among others.

Local spectral clustering techinques have been practically very successful (Leskovec et al., 2010; Andersen et al., 2012; Gleich & Seshadhri, 2012; Mahoney et al., 2012; Wu et al., 2012), which has led many authors to develop supporting theory (Spielman & Teng, 2013; Andersen & Peres, 2009; Gharan & 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 $\mathbf{X}$ 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]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]Other popular choices here include the unnormalized Laplacian, and symmetric normalized Laplacian.

1.1. Graph Connectivity Criteria

Here we define a pair of criteria that reflect the quality of a cluster with respect to $G = (V, E, w)$. There are many graph-based measures of cluster quality that one could consider; see, e.g., Yang & Leskovec (2015); Fortunato (2010) for an overview. The pair of criteria that we focus on are (arguably) quite natural, and moreover, they play a fundamental role in our analysis of the PPR algorithm. Our two criteria capture the *external* and *internal* connectivity of a subset $S \subseteq V$, denoted $\Phi(S; G)$ and $\Psi(S; G)$, respectively, and defined below in turn.

External Connectivity: Normalized Cut Define the cut between subsets $S, S' \subseteq V$ to be

$$\mathrm{cut}(S, S'; G) = \sum_{u \in S} \sum_{v \in S'} w_{uv},$$

and define $\mathrm{vol}(S; G) = \mathrm{cut}(S, V; G) = \sum_{u\in S} \sum_{v \in V} w_{uv}$. As our notion of external connectivity, we use the *normalized cut* of S, defined as

$$\Phi(S; G) = \frac{\mathrm{cut}(S; G)}{\min\{\mathrm{vol}(S; G), \mathrm{vol}(S^c; G)\}}, \tag{1}$$

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

Internal Connectivity: 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 $\mathbf{A}_S, \mathbf{D}_S$ be the adjacency matrix and degree matrix, respectively, of $G[S]$. Define the random walk matrix as usual, $\mathbf{W} = \mathbf{D}_S^{-1}\mathbf{A}_S$, and for $v \in V$, write

$$q_{vu}^{(t)} = e_v \mathbf{W}_S^t e_u$$

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

Our internal connectivity parameter will capture the time it takes for the random walk over $G[S]$ to mix (approach the stationary distribution) uniformly over S. For this, we first define *relative pointwise mixing time* of $G[S]$ as

$$\tau_\infty(G[S]) = \min\left\{ m : \frac{|q_{vu}^{(m)} - \widetilde{\pi}_u|}{\widetilde{\pi}_u} \le \frac{1}{4}, \text{ for } u, v \in V \right\},$$

Now our internal connectivity parameter is simply the inverse mixing time,

$$\Psi(S; G) = \frac{1}{\tau_\infty(G[S])}. \tag{2}$$

[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.

If S has normalized cut no greater than Φ, and inverse mixing time no less than Ψ, we call 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 approximate the optimal (Φ, Ψ)-cluster (or partition) for a given graph G.[3]

1.2. 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 $\mathbf{X}$: for a radius $r > 0$, we consider the r-*neighborhood graph* of $\mathbf{X}$, denoted $G_{n,r} = (V, E)$, an unweighted graph with vertices $V = \{1, \dots, n\}$, and an edge $(u, v) \in E$ if and only if $\|x_u - x_v\| \le r$, where $\|\cdot\|$ denotes Euclidean norm. Note that this is a special case of the general construction introduced above, with $K(u, v) = 1(\|x_u - x_v\| \le r)$.

Next, we define the PPR vector $\mathbf{p} = \mathbf{p}(v, \alpha; G_{n,r})$, 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:

$$\mathbf{p} = \alpha \mathbf{e}_v + (1 - \alpha)\mathbf{p}\mathbf{W}, \tag{3}$$

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 stationary measure $\pi_0 > 0$, we define a β-*sweep cut* of $\mathbf{p}$ as

$$S_\beta = \{u \in V : p_u > \beta \pi_0\}. \tag{4}$$

Having computed sweep cuts over a range $\beta \in (\frac{3}{10}, \frac{1}{2})$,[4] we output a cluster $\widehat{C} = S_{\beta^*}$, based on the sweep cut S_{β^*} that minimizes the normalized cut $\Phi(S_{\beta^*}; G_{n,r})$ as defined in (1). For concreteness, we summarize this procedure in Algorithm 1.

1.3. 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 $\mathcal{C}[\mathbf{X}] = \mathcal{C} \cap \mathbf{X}$ the *empirical density cluster*. Below we define a notion of consistency in density cluster estimation.

[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 in their paper 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.

[4]The choice of a specific range such as $(\frac{3}{10}, \frac{1}{2})$ is standard in the analysis of PPR algorithms, see, e.g., Zhu et al. (2013).

Algorithm 1 PPR on a Neighborhood Graph

Input: data $\mathbf{X} = \{x_1, \ldots, x_n\}$, radius $r > 0$, teleportation parameter $\alpha \in [0, 1]$, seed $v \in \mathbf{X}$, target stationary measure $\pi_0 > 0$.
Output: cluster $\widehat{C} \subseteq V$.

1: Form the neighborhood graph $G_{n,r}$.
2: Compute the PPR vector $\mathbf{p}(v, \alpha; G_{n,r})$ as in (3).
3: For $\beta \in (\frac{3}{10}, \frac{1}{2})$ compute sweep cuts S_β as in (4).
4: Return $\widehat{C} = S_{\beta^*}$, where

$$\beta^* = \underset{\beta \in (\frac{3}{10}, \frac{1}{2})}{\arg\min}\, \Phi(S_\beta; G_{n,r}).$$

Definition 1 (Consistent density cluster estimation). *For an estimator $\widehat{\mathcal{C}} \subseteq \mathbf{X}$ and cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$, we say $\widehat{\mathcal{C}}$ 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 $n \to \infty$:*

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

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 derive in Theorem 1 an upper bound on the normalized cut of a (thickened) empirical density cluster $\mathcal{C}_\sigma[\mathbf{X}]$, under natural geometric conditions (precluding clusters that are too thin and long).
2. Under largely the same set of geometric conditions, we derive in Theorem 2 a lower bound on the inverse mixing time of a random walk over $\mathcal{C}_\sigma[\mathbf{X}]$. We also provide in Theorem 3 a tighter lower bound, but under more restrictive assumptions (convexity of $\mathcal{C}_\sigma$).
3. In Section 3, we show in Theorem 4 that these bounds in bounds in Theorems 1 and 2, on the cluster quality criteria, have algorithmic consequences for PPR: properly initialized, Algorithm 1 performs consistent density cluster estimation in the sense of (5).
4. We show in Corollary 1 that Theorems 1 and 2, along with the results in Zhu et al. (2013), lead to alternative, graph-theoretic guarantees on cluster quality: an upper bound on the normalized cut of the estimated cluster $\widehat{C}$, and an upper bound on volume of the symmetric set difference between $\widehat{C}$ and $\mathcal{C}[\mathbf{X}]$.
5. In Section 4, we empirically demonstrate that violations of the geometric conditions we require manifestly impact density cluster recovery.

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.

1.4. 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, e.g., Tolliver & Miller 2006; von Luxburg 2007). In a similar vein, several recent works (McSherry, 2001; Rohe et al., 2011; Chaudhuri et al., 2012; Balakrishnan et al., 2011; Lei & Rinaldo, 2015; Abbe, 2018) have studied the efficacy of spectral methods in successfully recovering the community structure in the stochastic block model and variants.

Building on earlier work of Koltchinskii & 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.

Perhaps most similar to our results are the works Vempala & 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. Cluster Quality Critera Bounds for Density Clusters

2.1. Geometric Conditions on Density Clusters

In order to provide meaningful bounds on the normalized cut and inverse mixing time of an empirical density cluster, we must introduce conditions on the density f. Let $B(x, r) = \{y \in \mathbb{R}^d : \|y - x\| \leq r\}$ be the closed ball of radius $r \geq 0$, centered at $x \in \mathbb{R}^d$. Given a set $\mathcal{A} \subseteq \mathbb{R}^d$ and $\sigma > 0$, define $\mathcal{A}_\sigma = \mathcal{A} + B(0, \sigma) = \{y \in \mathbb{R}^d : \inf_{x \in \mathcal{A}} \|y - x\| \leq \sigma\}$, which we call the σ-expansion of $\mathcal{A}$.

We are now ready to give the 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

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

(A2) *Low noise density:* There exists $\gamma > 0$ such that for all $x \in \mathbb{R}^d$ with $0 < \mathrm{dist}(x, \mathcal{C}_\sigma) \le \sigma$,

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

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

(A3) *Cluster separation:* For all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$,

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

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

(A4) *Cluster diameter:* There exists $D < \infty$ such that for all $x, x' \in \mathcal{C}_\sigma$,

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

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

Assumptions (A1), (A2), and (A3) are used to upper bound $\Phi(\mathcal{C}[\mathbf{X}]; G_{n,r})$, whereas (A1) and (A4) are required to lower bound $\Psi(\mathcal{C}[\mathbf{X}]; G_{n,r})$. We note that the lower bound on minimum density in (A1) along with (A3) are similar to the (σ, ϵ)-saliency of Chaudhuri & Dasgupta (2010), a standard density clustering assumption, while (A2) is seen in, e.g., 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.

In the next several subsections, we will derive bounds on the cluster quality criteria evaluated on (σ-expansions of) density clusters. For notational simplicity, hereafter for $S \subseteq V$, we will abbreviate $\Phi(S; G_{n,r})$ by $\Phi_{n,r}(S)$, and similarly, $\Psi(S; G_{n,r})$ by $\Psi_{n,r}(S)$, and $\tau_\infty(G_{n,r}[S])$ by $\tau_{n,r}(S)$. We will also use ν for Lebesgue measure on $\mathbb{R}^d$, and $\nu_d = \nu(B)$ for the measure of the unit ball $B = B(0, 1)$.

2.2. Upper Bound on Normalized Cut

We start with an upper bound on the normalized cut (1) of $\mathcal{C}_\sigma[\mathbf{X}]$. (In the theorem, 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.)

Theorem 1. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1), (A2), and (A3), for some $\sigma, \lambda_\sigma, c_1, \gamma > 0$. Moreover, let*

$$\nu_{\mathbb{P}}(\mathcal{C}_\sigma) \le \frac{1}{2} \tag{6}$$

where $\nu_{\mathbb{P}}(\mathcal{A}) = \int_{\mathcal{A}} f(x) dx$ for $\mathcal{A} \subset \mathbb{R}^d$. Then for any $0 < r < \sigma/(4d)$, $0 < \delta < 1$, $\epsilon > 0$, and

$$n \ge \frac{9 \log(2/\delta)}{\epsilon^2} \left(\frac{1}{\lambda_\sigma^2 \nu(\mathcal{C}_\sigma) \nu_d r^d} \right)^2, \tag{7}$$

we have

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])}{r} \le 4 c_\sigma d \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon, \tag{8}$$

with probability at least $1 - \delta$, where $c_\sigma = 1/\sigma$.

Remark 1. The proof of Theorem 1, along with all other proofs in this paper, can be found in the supplementary document. The key idea is that for any $x \in \mathcal{C}$, the simple (possibly loose) fact $B(x, \sigma) \subseteq \mathcal{C}_\sigma$ translates into the upper bound $\nu(\mathcal{C}_{\sigma+r}) \le (1 + 2dr/\sigma)\nu(\mathcal{C}_\sigma)$. We leverage (A2) to find a corresponding bound on the weighted volume, then apply standard concentration inequalities to convert from population- to sample-based results.

Remark 2. The inequality in (8) is almost tight. Specifically, choosing $\mathcal{A}_\sigma = B(0, \sigma)$ and

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

we have that for n on the order of the lower bound in (7),

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

with probability at least $1 - \delta$, for some constant c. (Note that the factor of $1/\sigma$ in c_σ is not replicated above.)

2.3. Lower Bound on Inverse Mixing Time

Next we lower bound the inverse mixing time (2) of $\mathcal{C}_\sigma[\mathbf{X}]$, or equivalently, as $\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) = 1/\tau_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])$, we upper bound the mixing time.

Theorem 2. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1) and (A4) for some $\sigma, \lambda_\sigma, \Lambda_\sigma, D > 0$. Then for any $0 < r < \sigma/(4d)$, $0 < \delta < 1$, $\epsilon > 0$, and n satisfying*

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

where $\mu = \log(\frac{2D}{r})$, we have

$$\tau_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) \le (d \log \mu + c_\lambda) \cdot \left(c_1 + 9^d c_2 \frac{\Lambda_\sigma^4 D^{2d}}{\lambda_\sigma^4 r^{2d}} (d \log \mu + c_\lambda) \right), \tag{9}$$

with probability at least $1 - \delta$, where $c_1, c_2 > 0$ are universal constants 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 $G_{n,r}[\mathcal{C}_\sigma[\mathbf{X}]]$,

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

The factor of $1/r^{2d}$ in the bound in (9) is suboptimal. This exponential dependence on d stems from a loose bound on the aforementioned conductance. In particular, we assert only that any set $S \subseteq \mathcal{C}_\sigma[X]$ must have $\mathrm{cut}(S; \mathcal{C}_\sigma[\mathbf{X}])$ on the order of $n^2 r^{2d}$, while upper bounding $\mathrm{vol}(S; \mathcal{C}_\sigma[\mathbf{X}])$ 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 being a variant of classic results on rapid mixing (Jerrum & Sinclair, 1989).

2.4. Tighter Lower Bound Under Convexity

It is possible to sharpen the dependency on d in (9), but at the cost of an additional assumption.

(A5) *Convexity:* The set $\mathcal{C}_\sigma$ is convex.

With (A5) in place, we give a tighter bound on the mixing time (albeit one that holds only asymptotically).

Theorem 3. *Assume the conditions of Theorem 2, and additionally, (A5). Then for any $0 < r < \sigma/4d$, the following holds with probability 1:*

$$\begin{aligned}\limsup_{n\to\infty} \tau_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) \leq \\ & c_1 \frac{\Lambda_\sigma^8}{\lambda_\sigma^8}(\log\mu + c_\lambda)\Big(c_\lambda(d^3\log\mu + c_2) + \\ & \frac{d^2D^2}{r^2}\big(\log d + c_2\frac{\Lambda_\sigma^2}{\lambda_\sigma^2}c_\lambda + \\ & c_3 + \log\log\mu\big)\Big) + c_d\frac{o_r(1)}{r^2}, \end{aligned} \tag{10}$$

where $\mu = \log(\frac{2D}{r})$, $c_1, c_2, c_3 > 0$ are universal constants, c_d is constant in r but depends on dimension d, and $c_\lambda = \log(\Lambda_\sigma^2/\lambda_\sigma^2)$.

Remark 4. The only potentially exponential dependence on dimension comes from the factor of c_d. However, for sufficiently small values of r $o_r(1)/r^2$ will be dominated by $1/r^2$, and therefore the last term in (10) will contribute negligibly to the overall bound.

Remark 5. 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*,

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

where $\widetilde{\pi}$ 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

$$\int \frac{1}{t\widetilde{\Phi}_n(t)^2}dt, \tag{11}$$

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

Our proof relies on a novel (to the best of our knowledge) lower bound on this conductance function over $G_{n,r}$ in terms of a population-level analogue, which we denote $\widetilde{\Phi}_{\mathbb{P},r}$, and define formally in the supplementary document.

Lemma 1. *Fix $0 < t < 1/2$. Under the conditions on $\mathcal{C}_\sigma$ given by Theorem 3, the following statement holds: with probability one, as $n \to \infty$,*

$$\liminf_{n\to\infty} \widetilde{\Phi}_{n,r}(t) \geq \min\left\{\widetilde{\Phi}_{\mathbb{P},r}(t), c_d r\omega_r(1)\right\} \tag{12}$$

where c_d is the same as in Theorem 3.

Plugging $\widetilde{\Phi}_{\mathbb{P},r}$ into (11) yields a mixing time bound (with respect to total variation distance) on the order $dD^2/r^2 + d^2\log(D/r)$ over convex sets. By contrast, (10) 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 condition is only needed to lower bound $\widetilde{\Phi}_{\mathbb{P},r}$. Any lower bound on $\widetilde{\Phi}_{\mathbb{P},r}$ could be directly plugged into the machinery of the proof of Theorem 3 to yield an alternative result. In recent work, Abbasi-Yadkori et al. (2017) developed new population-level bounds on the conductance function (without requiring convexity), however the Markov chain dealt with there is somewhat different than the one we consider.

3. Consistent Cluster Estimation

3.1. Well-Conditioned Density Clusters

For PPR to successfully recover density clusters, the ratio $\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])/\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])$ should be small. Let us introduce the following notation for the upper bounds in Theorems 1 and 2:

$$\begin{aligned}\mathbf{\Phi}(\sigma,\lambda,\lambda_\sigma,\gamma) &= \frac{\lambda}{\lambda_\sigma}\frac{(\lambda_\sigma - \frac{\sigma^\gamma}{\gamma+1})}{\lambda_\sigma}, \\ 1/\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D) &= (d\log\mu + c_\lambda)\cdot \\ & \left(c_1 + 9^d c_2\frac{\Lambda_\sigma^4 D^{2d}}{\lambda_\sigma^4 r^{2d}}(d\log\mu + c_\lambda)\right)\end{aligned}$$

(where all constants $\lambda_\sigma, c_1, c_2, c_\lambda > 0$ are as in these theorems).

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 $\Lambda_\sigma, D < \infty$, and additionally let $\mathcal{C}_\sigma$ satisfy* (6). *Then, setting $\kappa_1(\mathcal{C})$ and $\kappa_2(\mathcal{C})$ to be*

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

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 (9), 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 (10), with no other changes to subsequent results.

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

$$\alpha \in [1/10, 1/9] \cdot \mathbf{\Psi}(\sigma, \lambda_\sigma, \Lambda_\sigma, D),\ r \leq \sigma/4d$$
$$\pi_0 \in [2/3, 6/5]\frac{\lambda_\sigma}{\nu(\mathcal{C}_\sigma)\Lambda_\sigma^2},\ v \in \mathcal{C}_\sigma[\mathbf{X}]^g \tag{13}$$

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* (13) *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*

$$\kappa_2 \leq \frac{1}{40 \cdot 36}\frac{\lambda_\sigma^2}{\Lambda_\sigma^2}\frac{r^d \nu_d}{\nu(\mathcal{C}_\sigma)}, \tag{14}$$

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

Remark 8. We note that replacing $40 \cdot 36$ by larger constants in (14) allow for wider ranges of parameters in (13).

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 of $\widehat{C}$ and the (volume-weighted) symmetric set difference of $\widehat{C}$ and $\mathcal{C}_\sigma[\mathbf{X}]$.

We collect some of the main results of (Zhu et al., 2013) in Theorem 5. For $G = (V, E)$ consider some $A \subseteq V$, and let $\Phi(A; G)$ and $\Psi(A; G)$ be defined as in (1) and (2), respectively.

Theorem 5 (Theorem 1 of (Zhu et al., 2013)). *There exists a set $A^g \subseteq A$ 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:

$$\mathrm{vol}(\widehat{C} \setminus A; G) \leq \frac{24\Phi(A; G)}{\Psi(A; G)}\mathrm{vol}(A)$$
$$\mathrm{vol}(A \setminus \widehat{C}; G) \leq \frac{30\Phi(A; G)}{\Psi(A; G)}\mathrm{vol}(A)$$
$$\Phi(\widehat{C}; G) = O\left(\frac{\Phi(A; G)}{\sqrt{\Psi(A; G)}}\right)$$

Corollary 1 – an analogous statement to Theorem 5 but stated specifically with respect to $\mathcal{C}_\sigma[\mathbf{X}]$– 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 (13)), the following guarantees hold for output set $\widehat{C} \subseteq \mathbf{X}$:*

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

4. Empirical Performance of PPR on Gaussian Mixture Models

The assumptions and theory of Section 2 – and therefore Section 3 as well– are tailored towards density functions with sharp transitions in gradient around the perimeter of the density cluster. These type of functions will satisfy (A1)

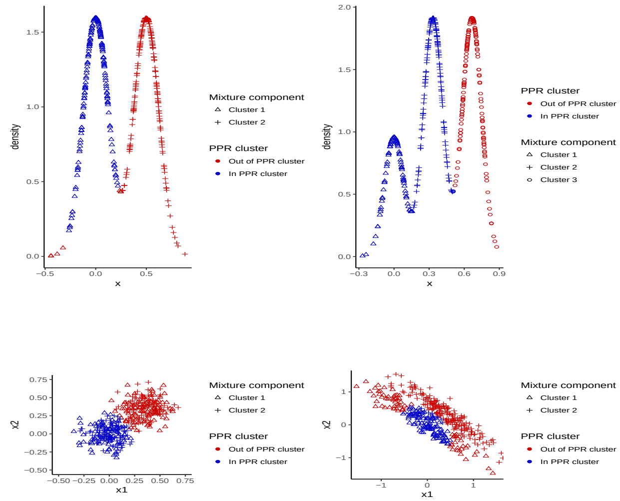

Figure 1. Algorithm 1 run over 4 example datasets. Parameters α and π_0 were tuned to minimize normalized cut of output set. Colors represent PPR cuts; shapes correspond to the maximum density among mixture components. Sample size $n = 400$.

with a small ratio of Λ_σ to λ_σ, while still satisfying (A2), potentially for $\gamma << 1$.

As mentioned previously, one line of work on spectral algorithms assesses their performance on mixture models. Particularly well-developed is the characterization of clustering performance on Gaussian Mixture Models (GMMs). Gaussian Mixture Models have smooth derivatives, and are therefore not good candidates for our work (at least, for reasonable values of the parameters in (A1) - (A4)). Since they are a classical choice in the clustering literature, we investigate PPR performance on them empirically.

Figure 1 shows data generated from four different Gaussian mixture models. The story follows roughly the same narrative as our theoretical work. The left two plots demonstrate instances, in one and two-dimensions, respectively, where the PPR cut has small error relative to a density cut. The right two plots, by contrast, show PPR failing to recover, even approximately, a density cut.

In the top right panel, the leftmost density cluster has a smaller cutsize than the set PPR ultimately recovers. However, the recovered set (colored in blue) has a higher volume, and therefore a lower normalized cut.

The bottom right panel shows the effect of a high diameter on the performance of PPR . Although any one vertex on the boundary of a density level set may have few cut edges, the length of the boundary is sufficiently long to make the entire cut too large for PPR to select it. Instead, the PPR cut defaults to a more geometrically compact shape, which has much smaller (unweighted) perimeter.

We also wish to investigate the robustness of these results to clustering problems with differing levels of difficulty. Consider the following mean vector and covariance matrices

$$\begin{aligned}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &= \big((0,0), (\theta,\theta)\big) \\ \boldsymbol{\Sigma}^{(1)} &= \begin{bmatrix} 1/64 & 0 \\ 0 & 1/64 \end{bmatrix} \\ \boldsymbol{\Sigma}^{(2)} &= \begin{bmatrix} 9/128 & -7/128 \\ -7/128 & 9/128 \end{bmatrix}.\end{aligned}$$

For a given choice of θ, let $f_1^{(\theta)}$ and $f_2^{(\theta)}$ be two-component Gaussian mixture densities supported on $\mathbb{R}^2$, defined as

$$\begin{aligned} f_1^{(\theta)}(x) &= \frac{1}{2}\phi_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}^{(1)}}(x) + \frac{1}{2}\phi_{\boldsymbol{\mu}_2;\boldsymbol{\Sigma}^{(1)}}(x) \quad \text{(isotropic model)} \\ f_2^{(\theta)}(x) &= \frac{1}{2}\phi_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}^{(2)}}(x) + \frac{1}{2}\phi_{\boldsymbol{\mu}_2,\boldsymbol{\Sigma}^{(2)}}(x) \quad \text{(anisotropic model)} \end{aligned}$$

where $\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(x)$ is the bivariate Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, evaluated at x.

For a given value of θ we can compute the symmetric set difference between the output of Algorithm 1, initialized near the origin, and the *oracle clusters*

$$\begin{aligned} C_1^{(\theta)} &= \{x : \phi_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}^{(1)}}(x) \geq \phi_{\boldsymbol{\mu}_2;\boldsymbol{\Sigma}^{(1)}}(x)\} \\ C_2^{(\theta)} &= \{x : \phi_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}^{(2)}}(x) \geq \phi_{\boldsymbol{\mu}_2;\boldsymbol{\Sigma}^{(2)}}(x)\} \end{aligned}$$

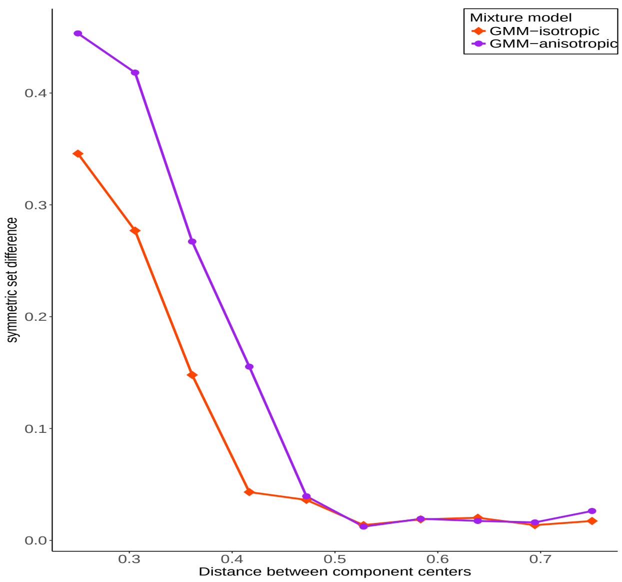

Figure 2. Symmetric set difference of Algorithm 1 run over GMM models with isotropic and anistropic component densities, as the distance between cluster centers changes.

(While these oracle clusters are not precisely density clusters as we have defined them, they exhibit similar behavior.) For a sequence of θ, we generate data from, for instance, density $f_i^{(\theta)}$, run Algorithm 1 (with hyperparameters tuned for lowest normalized cut) with output set $\widehat{C}_i$, and compute the symmetric set difference

$$\Delta(\widehat{C}_i, C_i^{(\theta)}) = \max\left\{\frac{\left|C_i^{(\theta)} \setminus \widehat{C}_i\right|}{\left|C_i^{\theta}\right|}, \frac{\left|\widehat{C}_i \setminus C_i^{\theta}\right|}{|\widehat{C}_i|}\right\}.$$

Figure 2 plots the relationship between $\sqrt{2}\theta = \|\boldsymbol{\mu_2} - \boldsymbol{\mu_1}\|_2$ and the symmetric set difference $\Delta(\widehat{C}_i, C_i^{(\theta)})$ for $i = 1, 2$. As we can see, for sufficiently large values of θ, PPR agrees with the oracle cluster in both the isotropic and anisotropic GMM models, while for sufficiently small values of θ it fails to even approximately recover the oracle cluster in either case. Throughout, the performance is superior for the isotropic model as opposed to the anisotropic one, as we would expect.

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 Figures 1 and 2 that poorly conditioned density clusters are not consistently estimated by PPR .

The initial motivation for this article was based on the intuition that density level sets, in the right conditions, will have small normalized cut. As a result, algorithms with normalized cut based guarantees (such as PPR) seemed likely to have density cluster recovery guarantees as well. However, the second-order behavior of PPR when failing to recover the conductance cut is also of interest. Are there situations when the conductance cut and density cut differ, yet PPR still recovers the latter? This is an open question.

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

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.

Andersen, R., Gleich, D. F., and Mirrokni, V. Overlapping clusters for distributed computation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 273–282. ACM, 2012.

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.

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.

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

Gleich, D. F. and Seshadhri, C. 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*, pp. 597–605. ACM, 2012.

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.

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.

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.

Wu, X., Li, Z., So, A. M., Wright, J., and fu Chang, S. Learning with partially absorbing random walks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 3077–3085. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4833-learning-with-partially-absorbing-random-walks.pdf`.

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.