# 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. 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. Finally, we 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 (**??**). This has led to recent increased interest in local spectral algorithms,

[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.

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 (**?**), and further developed by (**?????**), among others.

Local spectral clustering techinques have been practically very successful (**?????**), which has led many authors to develop supporting theory (**????**) 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 **?**).

### 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., (**??**) 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** For subsets $S, S' \subseteq V$, we define the cut, degree, and volume functionals as usual,

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

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 (lazy) random walk matrix as usual, $\mathbf{W} = \frac{\mathbf{D}_S^{-1}\mathbf{A}_S + \mathbf{I}_{|S|\times|S|}}{2}$, and for $v \in V$, 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 $v$.[2] 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 $\pi(u) = \deg(u; G[S])/\mathrm{vol}(S; G[S])$.)

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 the *relative pointwise mixing time* of $G[S]$ as

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

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 notation $e_v \mathbf{P}^t$ is used to denote the distribution of the random walk after $t$ steps.

If $S$ has normalized cut no greater than $\Phi$, and inverse mixing time no less than $\Psi$, we call it as a $(\Phi, \Psi)$-*cluster*. Both local (**?**) and global (**?**) spectral algorithms have been shown to output clusters (or partitions) which approximate the optimal $(\Phi, \Psi)$-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 $X$: 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:

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

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 $v$th 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 *$\beta$-sweep cut* of $p = (p_u)_{u \in V}$ as

$$S_\beta = \{u \in V : \frac{p_u}{\mathbf{D}_{uu}} > \frac{\beta}{\mathrm{vol}_0}\}. \tag{4}$$

Having computed sweep cuts over a range $\beta \in (\frac{1}{40}, \frac{1}{11})$,[4] we output a cluster $\widehat{C} = S_{\beta^*}$, based on the sweep cut $S_{\beta^*}$ that minimizes the normalized cut $\Phi(S_{\beta^*}; G_{n,r})$ as defined in (**??**). For concreteness, we summarize this procedure in Algorithm **??**.

**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 (**??**).
3: For $\beta \in (\frac{1}{40}, \frac{1}{11})$ compute sweep cuts $S_\beta$ as in (**??**).
4: Return $\widehat{C} = S_{\beta^*}$, where

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

### 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 $\mathcal{C}[X] = \mathcal{C} \cap X$ 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*

$$|\widehat{C} \setminus (\mathcal{S} \cap X)| + |(\mathcal{S} \cap X) \setminus \widehat{C}|. \tag{5}$$

**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 $n \to \infty$:*

$$\mathcal{C}[X] \subseteq \widehat{C} \quad \textit{and} \quad \widehat{C} \cap \mathcal{C}'[X] = \emptyset, \tag{6}$$

[3] In the case of (**?**), the internal connectivity parameter $\phi$ 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 $\phi^2 / \log(\mathrm{vol}(S)) \leq O(\Psi)$, and so the lower bound on $\phi$ translates to a lower bound on $\Psi$.

[4] 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., (**?**).

*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 **??**, we derive in Theorem **??** an upper bound on the normalized cut of a (thickened) empirical density cluster $\mathcal{C}_\sigma[X]$, under natural geometric conditions (precluding clusters that are arbitrarily thin).
2. Under additional geometric conditions, which exclude sets with large diameter or small bottleneck, we derive in Theorem **??** a lower bound on the inverse mixing time of a random walk over $\mathcal{C}_\sigma[X]$.
3. In Section **??**, we show in Theorems **??** and **??** that the bounds on the cluster quality criteria established in Theorems **??** and **??** have algorithmic consequences for PPR. Properly initialized, Algorithm **??** has low misclassification error with respect to a small enlargement of the set $\mathcal{C}$, and if the density cluster $\mathcal{C}$ is particularly well-conditioned, Algorithm **??** will perform consistent density cluster estimation in the sense of (**??**). Corollary **??** establishes that these statements hold also with respect to an approximate form of PPR, which can be efficiently computed.
4. In Section **??**, we empirically demonstrate the tightness of the bounds in Theorems **??** and **??**, 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.

### 1.4 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 (**??**) and their performance is well-understood in this context (see, e.g., (**??**)). In a similar vein, several recent works (**??????**) 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 (**?**), (**??**) 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 (**???**), 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 Criteria 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 $r > 0$, 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 $\sigma$-expansion of $\mathcal{A}$. For a differentiable function $g : \mathbb{R}^d \to \mathbb{R}^d$, write $\nabla g(x)$ to denote the Jacobian of $g$ evaluated at $x \in \mathbb{R}^d$.

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

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

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

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

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

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

(A4) *Lipschitz embedding:* There exists $\mathcal{K} \subseteq \mathbb{R}^d$ convex, and $g : \mathbb{R}^d \to \mathbb{R}^d$ satisfying, for some $L \geq 1$,

$$\det(\nabla g(x)) = 1, \frac{1}{L} \|x - y\| \leq \|g(x) - g(y)\| \leq L \|x - y\| \text{ for all } x, y \in \mathbb{R}^d$$

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 $x, x' \in \mathcal{K}$

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

Note that $\sigma$ plays several roles here, precluding arbitrarily narrow clusters in **??**, flat densities around the level set in **??**, and poorly separated clusters in **??**.

Assumptions **??**, **??**, and **??** are used to upper bound $\Phi(\mathcal{C}[X]; G_{n,r})$, whereas **??**, and **??** are required to lower bound $\Psi(\mathcal{C}[X]; G_{n,r})$. We note that the lower bound on minimum density in **??** along with **??** are similar to the $(\sigma, \epsilon)$-saliency of (**?**), a standard density clustering assumption, while **??** is seen in, e.g., (**?**) (as well as many other works on density clustering and level set estimation.) While **??** may be less standard, as we will see, it is critical in order to achieve reasonably tight bounds on $\Psi(\mathcal{C}[X]; G_{n,r})$. It is also worth highlighting that these assumptions are all local in nature, a benefit of studying a local algorithm such as PPR.

We emphasize that while many of these geometric conditions are typical in the density clustering literature, the restrictions we will impose upon them in order to obtain meaningful implications for PPR will not be. This is natural. The spectral algorithm we consider is not specifically designed for the task of level set estimation, and in fact one should expect PPR to fail to recover – either in the sense of (**??**), or indeed any reasonable notion of cluster recovery – a density cluster of sufficiently large diameter or sufficiently small thickness (though we do not provide any lower bounds to this effect). Indeed, one of the primary motivations of this work was to better understand and characterize the distinctions between those level sets which are well conditioned for spectral algorithms, and those which are not.

In the next several subsections, we will derive bounds on the cluster quality criteria evaluated on ($\sigma$-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 $\nu$ 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 (**??**) of $\mathcal{C}_\sigma[X]$. (In Theorem **??**, the upper bound on the density in Assumption **??** will not actually be needed, so we omit the parameter $\Lambda_\sigma > 0$ from the theorem statement.) For $\mathcal{S} \subseteq \mathbb{R}^d$ and $r > 0$, let

$$\pi_{\mathbb{P},r}(\mathcal{S}) := \frac{\int_{\mathcal{S}} \mathbb{P}(B(x, r)) f(x) dx}{\int_{\mathbb{R}^d} \mathbb{P}(B(x, r)) f(x) dx}.$$

**Theorem 1.** *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions **??**, **??**, and **??**, for some $\sigma, \lambda_\sigma, c_0, \gamma > 0$. Let $0 < r \leq \sigma/2d$ be such that*

$$\pi_{\mathbb{P},r}(\mathcal{C}_\sigma) \leq \frac{1}{2}. \tag{7}$$

*Then for any $0 < \delta < 1$, $\epsilon > 0$, if*

$$n \geq \frac{(2+\epsilon)^2 \log(3/\delta)}{\epsilon^2} \left( \frac{25}{6\lambda_\sigma^2 \nu(\mathcal{C}_\sigma)\nu_d r^d} \right)^2, \tag{8}$$

*then*

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[X])}{r} \leq c_1 \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon, \tag{9}$$

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

*Remark* 1. The proof of Theorem **??**, 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 fact $B(x, \sigma) \subseteq \mathcal{C}_\sigma$ translates into the upper bound $\nu(\mathcal{C}_\sigma + rB) \leq (1 + 2dr/\sigma)\nu(\mathcal{C}_\sigma)$. We leverage **??** 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 (**??**) is tight in the case of $\mathcal{C} = \{0\}$. To see this, let $\mathcal{C}_\sigma = B(0, \sigma)$ and

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

Then, some simple calculations yield

$$\mathbb{E}(\operatorname{cut}(\mathcal{C}_\sigma[X]; G_{n,r})) \geq c\lambda\nu_d r^d \mathbb{P}((\mathcal{C}_\sigma + B(0,r)), \quad \text{and} \quad \mathbb{E}(\operatorname{vol}_{n,r}(\mathcal{C}_\sigma[X]; G_{n,r}) \leq c'\lambda\nu_d r^d \mathbb{P}(\mathcal{C}_\sigma)$$

for some constants $c, c' > 0$. Thus the ratio $\mathbb{E}(\operatorname{cut}(\mathcal{C}_\sigma[X]; G_{n,r}))/\mathbb{E}(\operatorname{vol}_{n,r}(\mathcal{C}_\sigma[X]; G_{n,r})$ matches (**??**), up to constants.

### 2.3 Lower Bound on Inverse Mixing Time

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

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

$$\limsup_{n\to\infty} \tau_{n,r}(\mathcal{C}_\sigma[X]) \leq c_2 \frac{\Lambda_\sigma^4 d^3 D^2 L^2}{\lambda_\sigma^4 r^2} \log^2\left(\frac{1}{r}\right) + c_3 \log\left(\frac{\Lambda_\sigma}{\lambda_\sigma}\right) \tag{10}$$

for $c_2, c_3 > 0$ universal constants.

Our proof technique involves two key geometric quantities: the *local spread* $s(\widetilde{G}_{n,r})$ (where we abbreviate $\widetilde{G}_{n,r} := G_{n,r}[\mathcal{C}_\sigma[X]]$ and let $\widetilde{\pi}_{n,r}$ be the stationary distribution over $\widetilde{G}_{n,r}$) and the *conductance* $\widetilde{\Phi}_{n,r}$, defined respectively as

$$s(\widetilde{G}_{n,r}) := \frac{9}{10} \min_{x \in \mathcal{C}_\sigma[X]} \left\{ \deg(x; \widetilde{G}_{n,r}) \cdot \widetilde{\pi}_{n,r}(x) \right\}, \quad \widetilde{\Phi}_{n,r} = \min_{S \subseteq \mathcal{C}_\sigma[X]} \Phi(S; \widetilde{G}_{n,r}). \tag{11}$$

We argue that the random walk over $\widetilde{G}_{n,r}$ quickly escapes sets with stationary distribution less than $s(\widetilde{G}_{n,r})$, and so we avoid a $\log(1/\pi_0)$ 'start penalty' – where $\pi_0 := \min_{x \in \mathcal{C}_\sigma[X]} \widetilde{\pi}_{n,r} \lesssim \frac{1}{n}$ – characteristic to analyses of mixing time, which would render any resultant upper bound on mixing time vacuous. Instead, we obtain the tighter upper bound [5]

$$\tau_{n,r}(\mathcal{C}_\sigma[X]) \lesssim \frac{1}{\widetilde{\Phi}_{n,r}^2} \log^2\left(1/s(\widetilde{G}_{n,r})\right).$$

Then,

[5] For sequences $a_n, b_n$, we write $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) when there exists $c > 0$ such that $a_n \leq cb_n$ ($a_n \geq cb_n$) for all sufficiently large $n$.

- Some straightforward calculations yield $s(\widetilde{G}_{n,r}) \gtrsim \frac{\Lambda_\sigma r^d}{\lambda_\sigma}$.
- To handle the conductance, we introduce a continuous analogue,

$$\widetilde{\Phi}_{\mathbb{P},r} := \min_{\mathcal{S}\subseteq\mathcal{C}_\sigma} \left( \frac{\int_{\mathcal{S}} \mathbb{P}(B(x,r)\cap\mathcal{S}^c) f(x) dx}{\min\left\{\int_{\mathcal{S}} \mathbb{P}(B(x,r)\cap\mathcal{C}_\sigma) f(x) dx, \int_{\mathcal{S}^c} \mathbb{P}(B(x,r)\cap\mathcal{C}_\sigma) f(x) dx\right\}} \right), \quad \mathcal{S}^c = \mathcal{C}_\sigma \backslash \mathcal{S} \tag{12}$$

  and show the asymptotic lower bound $\limsup_{n\to\infty} \widetilde{\Phi}_{n,r} \gtrsim \widetilde{\Phi}_{\mathbb{P},r}$.
- Finally, we extend classical isoperimetric results lower bounding $\widetilde{\Phi}_{\mathbb{P},r}$, when $\mathbb{P}$ is uniform and $\mathcal{C}_\sigma$ convex, to hold under the more general conditions **??** and **??**.

*Remark* 3. The embedding assumption **??** and Lipschitz parameter $L$ obviously play an important role in the upper bound of Theorem **??**. It is clear that there is some interdependence between $L$ and other geometric parameters $\sigma$ and $D$, which might lead one to hope that **??** is non-essential. However, it is not possible to eliminate this condition without incurring an additional factor of at least $(D/\sigma)^d$ in (**??**), achieved, for instance, when $\mathcal{C}_\sigma$ consists of two balls of diameter $D$ linked by a cylinder of length $D$ and radius $\sigma$. (**??**) 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 **??** both satisfy **??** for reasonably small values of $L$.

# 3 Consistent Cluster Estimation

## 3.1 Well-Conditioned Density Clusters

For PPR to accurately estimate a set, the ratio of normalized cut to inverse mixing time should be small. Letting $\theta := (r, \sigma, \lambda, \lambda_\sigma, \Lambda_\sigma, \gamma, D, L)$ contain those parameters which govern the bounds given in Theorems **??** and **??**, further abbreviate

$$\mathbf{\Phi}(\theta) := c_1 r \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma}$$

$$\mathbf{\Psi}(\theta) := \left( c_2 \frac{\Lambda_\sigma^4 d^3 D^2 L^2}{\lambda_\sigma^4 r^2} \log^2\left(\frac{1}{r}\right) + c_3 \log\left(\frac{\Lambda_\sigma}{\lambda_\sigma}\right) \right)^{-1}$$

for these bounds (where all constants $c_0, c_1, c_2, c_3 > 0$ are as in these theorems).

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

**Definition 3** (Well-conditioned density clusters)**.** *For $\lambda > 0$ and $\mathcal{C} \in \mathbb{C}_f(\lambda)$, let $\mathcal{C}$ satisfy* **??** *-* **??** *for some $\sigma, \lambda, \lambda_\sigma, \Lambda_\sigma, \gamma, D, L > 0$, and additionally let $\mathcal{C}_\sigma$ satisfy* (**??**)*. Then, setting*

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

*we call $\mathcal{C}$ a $\kappa$-well-conditioned density cluster (with respect to $\theta$).*

We focus for a moment on the neighborhood graph radius $r$. While taking $r \to 0$ as $n \to \infty$—and thereby ensuring $G_{n,r}$ is sparse—is computationally attractive, the presence of a factor of $\frac{1}{r}$ 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 **??**, 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.

**Well-conditioned clusters.** 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 $\theta$), we require

$$r \leq \frac{\sigma}{2d}, \alpha \in [1/10, 1/9] \cdot \mathbf{\Psi}(\theta),$$
$$v \in \mathcal{C}_\sigma[X]^g, \mathrm{vol}_0 \in [3/4, 5/4] \cdot n(n-1) \int_{\mathcal{C}_\sigma} \mathbb{P}(B(x,r)) f(x) dx \tag{13}$$

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

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

In practice, a reasonable way to choose these hyperparameters is by tuning. For example, if one wanted to successfully recover a density cluster, one could vary each hyperparameter over a grid, retaining outputs $\widehat{C}$ of Algorithm **??** only if they recover some $\mathcal{C} \in \mathbb{C}_f(\lambda)$ and discarding them otherwise. Then simply return the minimum normalized cut set from those $\widehat{C}$ which were retained. Assuming there existed $\lambda > 0, \mathcal{C} \in \mathbb{C}_f(\lambda)$ such that $\mathcal{C}$ satisfied the conditions of Theorem **??**, and moreover some combination of tuning parameters in the chosen grid satisfied (**??**), this scheme would inherit the consistency guarantees of Theorem **??**.

**Misclassification error for** PPR**.** In **?**, building on the work of **?** and others, theory is developed which links algorithmic performance of PPR to the normalized cut and mixing time parameters. This work, combined with the results of Section **??**, immediately implies a bound on the volume of $\widehat{C} \setminus \mathcal{C}_\sigma[X]$ (and likewise $\mathcal{C}_\sigma[X] \setminus \widehat{C}$),

$$\mathrm{vol}_{n,r}(\widehat{C} \setminus \mathcal{C}_\sigma[X]), \mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X] \setminus \widehat{C}) \lesssim \kappa(\mathcal{C}) \mathrm{vol}_{n,r}(\mathcal{C}_\sigma[X]). \tag{14}$$

where we've written $\mathrm{vol}_{n,r}(S) := \mathrm{vol}(S; G_{n,r})$ for $S \subseteq X$. To translate (**??**) 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:* $0 < \lambda_{\min} < f(x)$ 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 **??** will be satisfied if, for instance, $\mathcal{X}$ is a $\sigma$-expansion.

**Theorem 3.** *Fix $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a $\kappa$-well conditioned density cluster (with respect to some $\theta$), and additionally assume $f$ satisfies* **??**. *Then, with probability tending to one as $n \to \infty$,*

$$\frac{|\mathcal{C}_\sigma[X] \setminus \widehat{C}|}{\left|\mathcal{C}_\sigma[X]\right|} \leq c_5 \kappa(\mathcal{C}) \frac{\Lambda_\sigma}{\lambda_\sigma}, \quad \textit{and} \quad \frac{|\widehat{C} \setminus \mathcal{C}_\sigma[X]|}{\left|\mathcal{C}_\sigma[X]\right|} \leq c_6 \kappa(\mathcal{C}) \frac{\Lambda_\sigma}{\lambda_{\min}}. \tag{15}$$

*for universal constants $c_4, c_5 > 0$.*

*Remark* 4. A notable implication of our theory is that as the diameter $D$ increases, our upper bound on the normalized cut of $\mathcal{C}_\sigma[X]$ remains unchanged, but $\kappa(\mathcal{C})$, and therefore the misclassification error, worsens (increases). This phenomenon reflects established wisdom regarding spectral partitioning algorithms more generally **??**, 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 $G_{n,r}$. This intuition will be supported by simulations in Section **??**.

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

$$\frac{p_{u'}}{\mathbf{D}_{uu}} \leq \frac{1}{40\mathrm{vol}_0} < \frac{1}{11\mathrm{vol}_0} \leq \frac{p_u}{\mathbf{D}_{uu}}. \tag{16}$$

so that any sweep cut $S_\beta$ for $\beta\mathrm{vol}_0 \in [1/40, 1/11]$ (i.e. any sweep cut considered by Algorithm **??**) will fulfill both conditions laid out in (**??**). In Theorem **??**, 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 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 $\mathcal{C}' \in \mathbb{C}_f(\lambda), \mathcal{C}' \neq \mathcal{C}$, for all $x \in \mathcal{C}' + \sigma B$,

$$\lambda_\sigma \leq f(x)$$

where $\sigma, \lambda_\sigma$ are as in **??**.

**Theorem 4.** *Fix $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ be a $\kappa$-well conditioned cluster (with respect to some $\theta$), and additionally assume* **??** *holds. If Algorithm* **??** *is well-initialized, there exists universal constant $c_7 > 0$ such that if*

$$\kappa(\mathcal{C}) \leq c_7 \frac{\lambda_\sigma^2 r^d \nu_d}{\Lambda_\sigma \mathbb{P}(\mathcal{C}_\sigma)}, \tag{17}$$

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

**Cluster estimation with the approximate** PPR **vector.** As mentioned previously, in practice exactly solving (**??**) may be too computationally expensive. To address this limitation, **?** introduced the *$\epsilon$-approximate* PPR *vector* (aPPR), which we will denote $p^{(\epsilon)}$. We refer the curious reader to **?** 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:

$$\text{for all } x \in X, \quad p(x) - \epsilon \deg_{n,r}(x) \leq p^{(\epsilon)}(x) \leq p(x) \tag{18}$$

Application of (**??**) within the proofs of Theorems **??** and **??** 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 $\kappa$-well-conditioned cluster (with respect to some $\theta$). Choose input parameters $\alpha, r, \mathrm{vol}_0, v$ to be well-initialized in the sense of* (**??**)*, set $\epsilon = \frac{1}{20\mathrm{vol}_0}$, and modify Algorithm* **??** *to compute the a*PPR *vector $p^{(\epsilon)}$ rather than the exact* PPR *vector $p$, with resulting output $\widehat{C}$.*

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

# 4 Experiments

## 4.1 Validating Theoretical Bounds

As we do not provide any theoretical lower bounds, we validate the tightness of Theorems **??** and **??** via simulation. We sample points according to the density function $q$, where for $x \in \mathbb{R}^d$

$$q(x) := \begin{cases} \lambda, & x \in [0, \sigma] \times D^{d-1} =: \mathcal{C}, \\ \lambda - \mathrm{dist}(x, \mathcal{C})\eta, & x \in \mathcal{C}_\sigma \setminus \mathcal{C}, \\ (\lambda - \sigma\eta) - \mathrm{dist}(x, \mathcal{C}_\sigma)^\gamma, & x \in (\mathcal{C}_\sigma + \sigma B) \setminus \mathcal{C}_\sigma, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $\lambda = \frac{150}{81}\sigma^\gamma$ and $\eta = \frac{15}{81}\sigma^{\gamma-1}$. Panels $(a)$ and $(b)$ in Figure **??** show $20,000$ samples from two parameterizations of $q$. In $(a)$, $\sigma = D = 3.2$, while in $(b)$ $\sigma = .1$ and $D = 3.2$. (For both, $d = 2$).

Panels $(c) - (f)$ in Figure **??** show the change in normalized cut and mixing time, respectively, as the parameters $\sigma$ ($(c)$ and $(d)$) and $D$ ($(e)$ and $(f)$) are varied. In panels $(c)$ and $(d)$ $\sigma = .1 \cdot \sqrt{2}^j, j =$

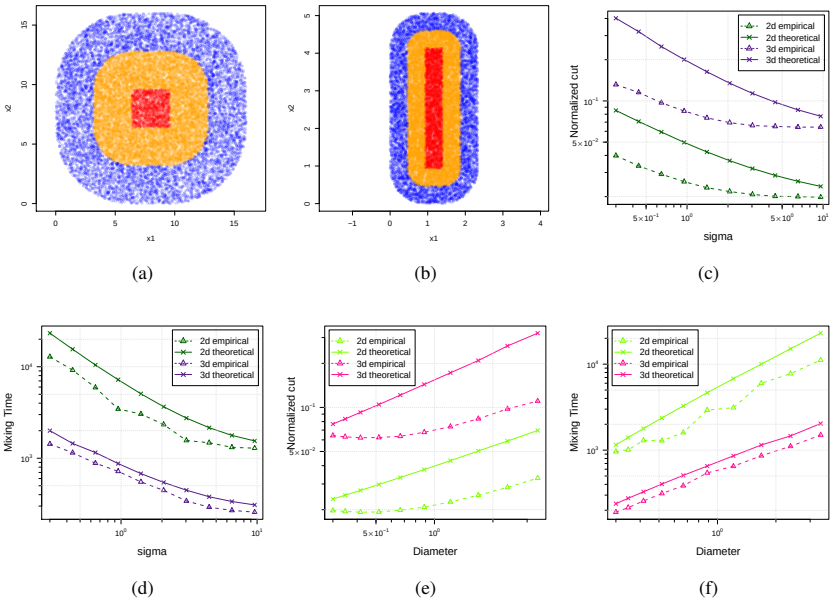


Figure 1: Samples, empirical results, and theoretical bounds for mixing time and normalized cut as diameter and thickness are varied. In (a) and (b), points in $\mathcal{C}$ are colored in red; points in $\mathcal{C}_\sigma \setminus \mathcal{C}$ are colored in yellow; and remaining points in blue.

$1, \ldots, 10$, and $D$ is fixed at 3.2. In panels $(e)$ and $(f)$, $D = .1 \cdot \sqrt{2}^j, j = 1, \ldots, 10$ and $\sigma$ is fixed at .1. For each panel, the solid lines show, up to constants [6], the theoretical upper bound, given by Theorem **??** for panels $(c)$ and $(e)$ and Theorem **??** for panels $(d)$ and $(f)$. The dashed lines show the computed empirical value, averaged over $m$ trials ($m = 100$ for the normalized cut, dashed lines in panels $(c)$ and $(e)$, and $m = 20$ for the mixing time, dashed lines in panel $(d)$ and $(f)$). For each trial across all parameters, $r$, the neighborhood graph radius, is set throughout to be as small as possible such that the resulting graph is connected, for computational efficiency. Green lines correspond to dimension $d = 2$, whereas purple/pink lines correspond to $d = 3$.

Panels $(d)$ and $(f)$ show the solid lines tracking closely to the dashed lines, in both 2 and 3 dimensions. This provides empirical evidence that the upper bound on mixing time given by Theorem **??** has the right dependency on both thickness parameter $\sigma$ 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 $\sigma \approx D$; this is the right hand side on panel $(c)$ and the left hand side on panel $(e)$. It is in these regions that the slopes of the dashed and solid lines are most different. As the ratio $D/\sigma$ grows, we see the slopes of the empirical curves becoming more similar to those predicted by theory. The takeaway message is that while the dependency in (**??**) on $\sigma$ and $D$ is loose for clusters with diameter close to thickness, it becomes tighter as $D/\sigma$ grows.

### 4.2 Empirical PPR, normalized cut, and density clustering comparison

To drive home the main implications of Theorems **??** and **??**, we show the behavior of PPR, normalized cut, and the density clustering algorithm of (**?**) on (a variant of) the famous 'two moons' dataset, considered a prototypical success story for spectral clustering algorithms. To form each of the four rows in Figure **??**, 800 points are independently sampled following a 'two moons plus

[6] 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.

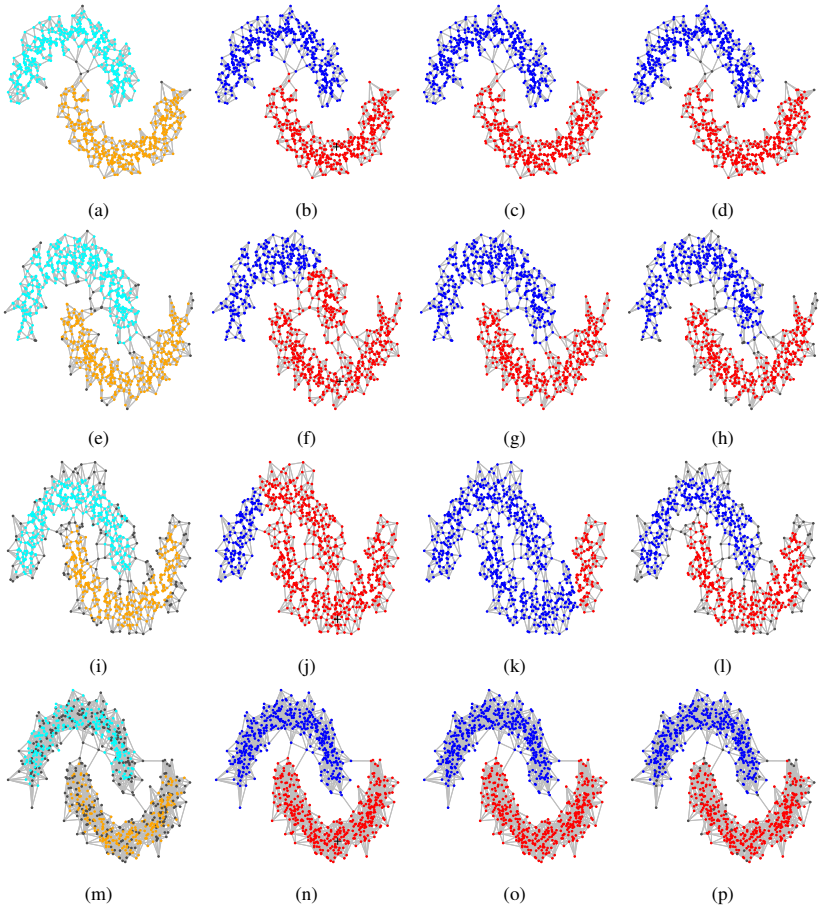


Figure 2: True density (column 1), PPR (column 2), normalized cut (column 3) and estimated density (column 4) clusters for 4 different simulated data sets. Seed node for PPR denoted by a black cross.

Gaussian noise model'. Formally, the (respective) generative models for the data are

$$Z \sim \text{Bern}(1/2), \theta \sim \text{Unif}(0, \pi) \tag{20}$$

$$X(Z, \theta) = \begin{cases} \mu_1 + (r\cos(\theta), r\sin(\theta)) + \sigma\epsilon, & \text{if } Z = 1 \\ \mu_2 + (r\cos(\theta), -r\sin(\theta)) + \sigma\epsilon, & \text{if } Z = 0 \end{cases} \tag{21}$$

where

$$\mu_1 = (-.5, 0),\ \mu_2 = (0, 0),\ \epsilon \sim N(0, I_2) \quad \text{(row 1)}$$
$$\mu_1 = (-.5, -.07),\ \mu_2 = (0, .07),\ \epsilon \sim N(0, I_2) \quad \text{(row 2)}$$
$$\mu_1 = (-.5, -.125),\ \mu_2 = (0, .125),\ \epsilon \sim N(0, I_2) \quad \text{(row 3)}$$
$$\mu_1 = (-.5, -.025),\ \mu_2 = (0, .025),\ \epsilon \sim N(0, I_{10}) \quad \text{(row 4)}$$

for $I_d$ the $d \times d$ identity matrix. The first column consists of the empirical density clusters $C_n$ and $C'_n$ for a particular threshold $\lambda$ of the density function; the second column shows the PPR plus minimum normalized sweep cut cluster, with hyperparameter $\alpha$ and all sweep cuts considered; the third column shows the global minimum normalized cut, computed according to the algorithm of **?**; and the last column shows a cut of the density cluster tree estimator of **?**.

Rows 1-3 show the degrading ability of PPR to recover density clusters as the two moons become less salient. In the first row, the normalized cut conforms to the density cluster, and PPR recovers both. In the second row, the normalized cut still conforms to the density cluster, but because the internal connectivity of the lower moon is low, PPR fails to recover the normalized cut. In the third row, the moons have such low saliency that even the normalized cut fails to recover the lower moon; we also

see from $(k)$ that PPR does not somehow save us in this situation. Note that this is not a function of the finite sample: the 4th column shows us that a well-designed density clustering algorithm can recover the true density cluster.

The fourth row illustrates the effect of dimension. The gray dots in $(m)$ (as in $(a)$, $(e)$ and $(i)$ are observations in low-density regions. While the PPR sweep cut $(n)$ has relatively high symmetric set difference with the chosen density cut, it still recovers $C_n$ in the sense of Definition **??**.

# 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.