

---

# Local Spectral Clustering of Density Upper Level Sets

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We analyze Personalized PageRank (PPR), a local spectral method for clustering,  
2 which extract clusters using locally-biased random walks around a user-specified  
3 seed node. In contrast to previous work, we adopt a traditional statistical learning  
4 setup, where we obtain samples from an unknown distribution, and aim to identify  
5 connected regions of high-density (density clusters). We prove that PPR, run on a  
6 neighborhood graph, extracts sufficiently salient density clusters. We also provide  
7 empirical support for our theory.

## 8 1 Introduction

9 In this paper, we study the problem of clustering: splitting a given data set into groups that satisfy  
10 some notion of within-group similarity and between-group difference. We focus on spectral clustering,  
11 a family of powerful nonparametric clustering algorithms. Generally speaking, a spectral algorithm  
12 first constructs a geometric graph  $G$ , where vertices correspond to samples, and edges correspond to  
13 proximities between samples. It then learns a feature embedding based on the Laplacian of  $G$ , and  
14 applies a simple clustering technique (like k-means clustering) in the embedded feature space.

15 When applied to geometric graphs built from a large number of samples, global spectral clustering  
16 methods can be computationally cumbersome and insensitive to the local geometry of the underlying  
17 distribution [Leskovec et al., 2010, Mahoney et al., 2012]. This has led to increased interest in local  
18 spectral clustering algorithms, which leverage locally-biased spectra computed using random walks  
19 around some user-specified seed node. A popular local clustering algorithm is Personalized PageRank  
20 (PPR), first introduced by Haveliwala [2003], then further developed by [Spielman and Teng, 2011,  
21 2014, Andersen et al., 2006, Mahoney et al., 2012, Zhu et al., 2013], among others.

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

31 **PPR on a neighborhood graph.** We now describe the clustering algorithm that will be our focus  
32 for the rest of the paper. Let  $X = \{x_1, \dots, x_n\}$  be a sample drawn i.i.d. from a distribution  $\mathbb{P}$  on  
33  $\mathbb{R}^d$ , with density  $f$ . For a radius  $r > 0$ , we define  $G_{n,r} = (V, E)$  to be the  $r$ -neighborhood graph  
34 of  $X$ , an unweighted, undirected graph with vertices  $V = X$ , and an edge  $(x_i, x_j) \in E$  if and only  
35 if  $\|x_i - x_j\| \leq r$ , where  $\|\cdot\|$  is the  $\ell_2$  norm. We denote by  $A \in \mathbb{R}^{n \times n}$  the adjacency matrix, with  
36 entries  $A_{uv} = 1$  if  $(u, v) \in E$  and 0 otherwise. We also denote by  $D$  the diagonal degree matrix,  
37 with  $D_{uu} = \sum_{v \in V} A_{uv}$ , and by  $I$  the  $n \times n$  identity matrix.

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

$$p = \alpha e_v + (1 - \alpha)pW, \quad (1)$$

40 where  $W = (I + D^{-1}A)/2$  is the lazy random walk matrix over  $G_{n,r}$  and  $e_v$  is the indicator vector  
 41 for node  $v$  (that has a 1 in the  $v$ th position and 0 elsewhere). For a level  $\beta > 0$  and a target volume  
 42  $\text{vol}_0 > 0$ , we define a  $\beta$ -sweep cut of  $p = (p_u)_{u \in V}$  as

$$S_\beta := \left\{ u \in V : \frac{p_u}{D_{uu}} > \frac{\beta}{\text{vol}_0} \right\}. \quad (2)$$

43 We will use the normalized cut metric to determine which sweep cut  $S_\beta$  is the best cluster estimate.  
 44 For a set  $S \subseteq V$  with complement  $S^c = V \setminus S$ , we define  $\text{cut}(S; G_{n,r}) := \sum_{u \in S, v \in S^c} A_{uv}$ , and  
 45  $\text{vol}(S; G_{n,r}) := \sum_{u \in S} D_{uu}$ . We define the *normalized cut* of  $S$  as

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

46 Having computed sweep cuts  $S_\beta$  over a range  $\beta \in (\frac{1}{40}, \frac{1}{11})^1$ , we output the cluster estimate  $\hat{C} = S_{\beta^*}$   
 47 that has minimum normalized cut. For concreteness, this is summarized in Algorithm 1.

---

**Algorithm 1** PPR on a neighborhood graph

---

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

**Output:** cluster  $\hat{C} \subseteq V$ .

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

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


---

48 **Estimation of density clusters.** Let  $\mathbb{C}_f(\lambda)$  denote the connected components of the density upper  
 49 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  
 50 *empirical density cluster*. The size of the symmetric set difference between estimated and empirical  
 51 clusters is a commonly used metric to quantify cluster estimation error [Korostelev and Tsybakov,  
 52 1993, Polonik, 1995, Rigollet and Vert, 2009].

53 **Definition 1** (Symmetric set difference). For an estimator  $\hat{C} \subseteq X$  and set  $\mathcal{S} \subseteq \mathbb{R}^d$ , we define

$$\Delta(\hat{C}, \mathcal{S}) := |\hat{C} \setminus \mathcal{S}[X] \cup \mathcal{S}[X] \setminus \hat{C}|, \quad (4)$$

54 the cardinality of the symmetric set difference between  $\hat{C}$  and  $\mathcal{S} \cap X = \mathcal{S}[X]$ .

55 However, the symmetric set difference does not measure whether  $\hat{C}$  can distinguish any two distinct  
 56 clusters  $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_f(\lambda)$ . We therefore also study a second notion of cluster estimation, first introduced  
 57 by Hartigan [1981], and defined asymptotically.

58 **Definition 2** (Consistent density cluster estimation). For an estimator  $\hat{C} \subseteq X$  and cluster  $\mathcal{C} \in \mathbb{C}_f(\lambda)$ ,  
 59 we say  $\hat{C}$  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  
 60  $n \rightarrow \infty$ :

$$\mathcal{C}[X] \subseteq \hat{C} \quad \text{and} \quad \hat{C} \cap \mathcal{C}'[X] = \emptyset, \quad (5)$$

61 with probability tending to 1.

---

<sup>1</sup>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].

62 **Summary of results.** A summary of our results (and outline for this paper) is as follows.

- 63 1. In Section 2, we introduce a set of natural geometric conditions on the density cluster  $\mathcal{C}$  and  
 64 show that when Algorithm 1 is properly initialized, the size of the symmetric set difference  
 65 of  $\mathcal{C}$  and a thickened version of the density cluster  $\mathcal{C}_\sigma$  can be bounded in a meaningful way.
- 66 2. We further show in Section 2 that if the density cluster  $\mathcal{C}$  is particularly well-conditioned,  
 67 Algorithm 1 will consistently estimate a density cluster in the sense of (5).
- 68 3. In Section 3, we detail some of the analysis required to prove our main results, and expose  
 69 the parts various geometric quantities play in the difficulty of the clustering problem.
- 70 4. In Section 4, we empirically investigate the tightness of our analysis, and provide examples  
 71 showing how violations of our geometric conditions impact density cluster recovery by PPR.

72 Our main takeaway: PPR, run on a neighborhood graph, recovers geometrically compact high-density  
 73 clusters.

74 **Related work.** In addition to the background on local spectral clustering given previously, a few  
 75 related lines of work are worth highlighting. [Shi et al., 2009, Schiebinger et al., 2015] examine the  
 76 consistency of spectral algorithms in recovering the latent labels in certain nonparametric mixture  
 77 models. Their results focus on global rather than local methods, and thus impose global rather than  
 78 local conditions on the nature of the density. Moreover, they do not in general guarantee recovery  
 79 of density clusters, which is the focus in our work. Perhaps most importantly, these works rely on  
 80 general cluster saliency conditions, which implicitly depend on many distinct geometric aspects of  
 81 the cluster  $\mathcal{C}$  under consideration. We make this dependence more explicit, and in doing so expose  
 82 the role each geometric condition plays in the clustering problem.

83 More broadly, density clustering and level set estimation is a well-studied problem. Polonik [1995],  
 84 Rigollet and Vert [2009] study density clustering under the symmetric set difference metric, Tsybakov  
 85 [1997], Singh et al. [2009] describe minimax optimal level set estimators under Hausdorff loss and  
 86 Hartigan [1981], Chaudhuri and Dasgupta [2010] consider consistent estimation of the cluster tree, to  
 87 note but a few works. Our goal is not to improve on these results, nor to offer a better algorithm for  
 88 level set estimation; indeed, seen as a density clustering algorithm, PPR has none of the optimality  
 89 guarantees found in the aforementioned works. Instead, our motivation is start with a widely-used  
 90 local spectral method, PPR, and to better understand and characterize the distinctions between those  
 91 density clusters which are well-conditioned for PPR, and those which are not.

## 92 2 Estimation of well-conditioned density clusters

93 We formalize some geometric conditions, and use these to define a condition number  $\kappa(\mathcal{C})$ , which  
 94 measures the difficulty PPR will have in estimating  $\mathcal{C}$ . (Our theoretical guarantees for PPR will be  
 95 framed in terms of  $\kappa(\mathcal{C})$ .)

96 **Geometric conditions on density clusters.** At a high level, for PPR to be successful, the underly-  
 97 ing density cluster must be geometrically well-conditioned. At a minimum, we want to avoid sets that  
 98 contain arbitrarily thin bridges or spikes. Hence, as in Chaudhuri and Dasgupta [2010], we consider  
 99 a thickened version of  $\mathcal{C} \in \mathbb{C}_f(\lambda)$  defined as  $\mathcal{C}_\sigma := \{x \in \mathbb{R}^d : \text{dist}(x, \mathcal{C}) \leq \sigma\}$ , which we call the  
 100  $\sigma$ -expansion of  $\mathcal{C}$ . Here  $\text{dist}(x, \mathcal{C}) := \inf_{y \in \mathcal{C}} \|y - x\|$ . We now list our conditions on  $\mathcal{C}_\sigma$ .

101 (A1) *Bounded density within cluster:* There exist constants  $0 < \lambda_\sigma < \Lambda_\sigma < \infty$  such that  
 102  $\lambda_\sigma \leq \inf_{x \in \mathcal{C}_\sigma} f(x) \leq \sup_{x \in \mathcal{C}_\sigma} f(x) \leq \Lambda_\sigma$ .

103 (A2) *Cluster separation:* For all clusters  $\mathcal{C}' \in \mathbb{C}_f(\lambda)$  with  $\mathcal{C}' \neq \mathcal{C}$ ,  $\text{dist}(\mathcal{C}_\sigma, \mathcal{C}'_\sigma) > \sigma$ , where  
 104  $\text{dist}(\mathcal{C}_\sigma, \mathcal{C}'_\sigma) := \inf_{x \in \mathcal{C}_\sigma} \text{dist}(x, \mathcal{C}'_\sigma)$ .

105 (A3) *Low noise density:* There exist  $\gamma, c_0 > 0$  such that for  $x \in \mathbb{R}^d$  with  $0 < \text{dist}(x, \mathcal{C}_\sigma) \leq \sigma$ ,  
 106  $\inf_{x' \in \mathcal{C}_\sigma} f(x') - f(x) \geq c_0 \text{dist}(x, \mathcal{C}_\sigma)^\gamma$ .

107 (A4) *Lipschitz embedding:* There exists  $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$  with the following properties: i) we have  
 108  $\mathcal{C}_\sigma = g(\mathcal{K})$ , for a convex set  $\mathcal{K} \subseteq \mathbb{R}^d$  with  $\text{diam}(\mathcal{K}) = \sup_{x, y \in \mathcal{K}} \|x - y\| =: \rho < \infty$ ; ii)

109  $\det(\nabla g(x)) = 1$  for all  $x \in \mathcal{C}_\sigma$ , where  $\nabla g(x)$  is the Jacobian of  $g$  evaluated at  $x$ ; and iii)  
 110 for some  $L \geq 1$ ,

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

111 Succinctly,  $\mathcal{C}_\sigma$  is the image of a convex set with finite diameter under a measure preserving,  
 112 bi-Lipschitz transformation.

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

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

114 where  $B(x, r)$  is the closed ball of radius  $r$  at  $x$ .

115 To motivate these conditions, Zhu et al. [2013] show for arbitrary graph  $G = (V, E)$  and subset of  
 116 vertices  $S \subseteq V$ , the PPR estimate  $\hat{C}$  of subset  $S$  satisfies, for a constant  $c > 0$ ,

$$\text{vol}(\hat{C} \setminus S; G) + \text{vol}(S \setminus \hat{C}; G) \leq c(\Phi(S, G) \cdot \tau_\infty(G[S])) \text{vol}(S; G), \quad (6)$$

117 where  $\Phi(S; G)$  is the normalized cut of  $S$  (as defined in (3)), and  $\tau_\infty(G[S])$  is called the *mixing*  
 118 *time* of a random walk over the induced subgraph  $G[S]$  (to be defined precisely later, in (16)). The  
 119 left-hand side in (6) resembles a (degree-weighted) form of the symmetric set difference metric in  
 120 (4). As we will show in Section 3, the conditions (A1)–(A5) allow us to upper bound the normalized  
 121 cut  $\Phi(\mathcal{C}_\sigma[X]; G_{n,r})$ , and the mixing time  $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$ . (Informally, (A2), (A3) yield an upper  
 122 bound on  $\text{cut}(\mathcal{C}_\sigma[X]; G_{n,r})$ , and (A1) yields a lower bound on  $\text{vol}(\mathcal{C}_\sigma[X]; G_{n,r})$ ; together with (A5),  
 123 this gives an upper bound on the normalized cut. On the other hand, (A1), (A4) preclude bottlenecks  
 124 in the induced subgraph  $G_{n,r}[\mathcal{C}_\sigma[X]]$ , and combined with (A4), this leads to an upper bound on the  
 125 mixing time over this subgraph.)

126 **Condition number.** Motivated by (6), we will define  $\kappa(\mathcal{C})$  to be an upper bound on the product  
 127  $\Phi(\mathcal{C}_\sigma[X]; G_{n,r}) \cdot \tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$ . The smaller  $\kappa(\mathcal{C})$  is, the more success PPR will have in recover-  
 128 ing  $\mathcal{C}$ . Let  $\theta := (r, \sigma, \lambda, \lambda_\sigma, \Lambda_\sigma, \gamma, \rho, L)$  contain the geometric parameters from (A1)–(A5).

129 **Definition 3** (Well-conditioned density clusters). *For  $\lambda > 0$  and  $\mathcal{C} \in \mathbb{C}_f(\lambda)$ , let  $\mathcal{C}$  satisfy (A1)–(A5)*  
 130 *for some  $\theta$ . Then, for universal constants  $c_1, c_2, c_3 > 0$  to be specified later, we set*

$$\Phi_u(\theta) := c_1 r \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma}, \quad \tau_u(\theta) := c_2 \frac{\Lambda_\sigma^4 d^3 \rho^2 L^2}{\lambda_\sigma^4 r^2} \log^2 \left( \frac{1}{r} \right) + c_3, \quad (7)$$

131 and letting  $\kappa(\mathcal{C}) := \Phi_u(\theta) \cdot \tau_u(\theta)$ , we call  $\mathcal{C}$  a  $\kappa$ -well-conditioned density cluster.

132 We note that  $\Phi_u(\theta)$  and  $\tau_u(\theta)$  are exactly the upper bounds on  $\Phi(\mathcal{C}_\sigma[X]; G_{n,r})$  and  $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$   
 133 that we derive in our analysis later, in Section 3.

134 **Well-initialized algorithm.** As is typical in the local clustering literature, our algorithmic results  
 135 will be stated with respect to specific ranges of each of the user-specified parameters. In particular,  
 136 for a well-conditioned density cluster  $\mathcal{C}$  (with respect to some  $\theta$ ), we require

$$0 < r \leq \frac{\sigma}{2d}, \quad \alpha \in [1/10, 1/9] \cdot \frac{1}{\tau_u(\theta)}, \quad (8)$$

$$v \in \mathcal{C}_\sigma[X]^g, \quad \text{vol}_0 \in [3/4, 5/4] \cdot n(n-1) \int_{\mathcal{C}_\sigma} \mathbb{P}(B(x, r)) f(x) dx,$$

137 where  $\mathcal{C}_\sigma[X]^g \subseteq \mathcal{C}_\sigma[X]$  will be some large (“good”) subset of  $\mathcal{C}_\sigma[X]$ . In particular, abbreviating  
 138  $\text{vol}_{n,r}(S) := \text{vol}(S; G_{n,r})$  for  $S \subseteq X$ , we will have  $\text{vol}_{n,r}(\mathcal{C}_\sigma[X]^g) \geq \text{vol}_{n,r}(\mathcal{C}_\sigma[X])/2$ .

139 **Definition 4.** *If the input parameters to Algorithm 1 satisfy (8) for some well-conditioned density*  
 140 *cluster  $\mathcal{C}$ , we say the algorithm is well-initialized.*

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

145 **Main theorems.** The results of Section 3, combined with (6), give an upper bound on the volume  
 146 of  $\widehat{C} \setminus \mathcal{C}_\sigma[X]$  and  $\mathcal{C}_\sigma[X] \setminus \widehat{C}$ ,

$$\text{vol}_{n,r}(\widehat{C} \setminus \mathcal{C}_\sigma[X]) + \text{vol}_{n,r}(\mathcal{C}_\sigma[X] \setminus \widehat{C}) \leq c\kappa(\mathcal{C})\text{vol}_{n,r}(\mathcal{C}_\sigma[X]). \quad (9)$$

147 To translate (9) into meaningful bounds on the symmetric set difference metric  $\Delta(\mathcal{C}_\sigma[X], \widehat{C})$ , we  
 148 want to preclude vertices  $x \in X$  from having arbitrarily small degree, and so we make some regularity  
 149 assumptions on  $\mathcal{X} := \text{supp}(f)$ . Let  $\nu$  denote the Lebesgue measure on  $\mathbb{R}^d$ , and  $\nu_d := \nu(B)$  be the  
 150 measure of the unit ball  $B = B(0, 1)$ .

151 (A6) *Regular support:* There exists some constant  $\lambda_{\min} > 0$  such that  $\lambda_{\min} < f(x)$  for all  $x \in \mathcal{X}$ .  
 152 Additionally, there exists some  $c > 0$  such that for each  $x \in \partial\mathcal{X}$ ,  $\nu(B(x, r) \cap \mathcal{X}) \geq c\nu_d r^d$ .

153 Note that the latter condition in (A6) will hold if the boundary  $\partial\mathcal{X}$  is sufficiently regular. Now we  
 154 present our main bound on the symmetric set difference metric.

155 **Theorem 1.** Fix  $\lambda > 0$ , let  $\mathcal{C} \in \mathbb{C}_f(\lambda)$  be a  $\kappa$ -well-conditioned density cluster (with respect to some  
 156  $\theta$ ), and additionally assume  $f$  satisfies (A6). If Algorithm 1 is well-initialized, there exists a universal  
 157 constant  $c_4 > 0$  such that with probability tending to 1 as  $n \rightarrow \infty$ ,

$$\Delta(\mathcal{C}_\sigma[X], \widehat{C}) \leq c_4\kappa(\mathcal{C}) \frac{\Lambda_\sigma}{\lambda_{\min}}. \quad (10)$$

158 The proof of Theorem 1, along with all other proofs in this paper, is deferred to the supplementary  
 159 material. Note that this result says the symmetric set difference metric  $\Delta(\mathcal{C}_\sigma[X], \widehat{C})$  is proportional  
 160 to the difficulty of the clustering problem, as measured by the condition number  $\kappa(\mathcal{C})$ .

161 Neither (9) nor (10) imply consistent density cluster estimation in the sense of (5). This notion of  
 162 consistency requires a uniform bound over  $p$ : for all  $\mathcal{C}' \in \mathbb{C}_f(\lambda)$ ,  $\mathcal{C}' \neq \mathcal{C}$ , and each  $u \in \mathcal{C}$ ,  $w \in \mathcal{C}'$ ,  
 163

$$\frac{p_w}{D_{ww}} \leq \frac{1}{40\text{vol}_0} < \frac{1}{11\text{vol}_0} \leq \frac{p_u}{D_{uu}}, \quad (11)$$

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

169 (A7) *Bounded density in other clusters:* Letting  $\sigma, \lambda_\sigma$  be as in (A1), for each  $\mathcal{C}' \in \mathbb{C}_f(\lambda)$  and for  
 170 all  $x \in \mathcal{C}'_\sigma$ ,  $\lambda_\sigma \leq f(x)$ .

171 Next we give our main result on consistent cluster recovery by PPR.

172 **Theorem 2.** Fix  $\lambda > 0$ , let  $\mathcal{C} \in \mathbb{C}_f(\lambda)$  be a  $\kappa$ -well-conditioned density cluster (with respect to some  
 173  $\theta$ ), and additionally assume  $f$  satisfies (A7). If Algorithm 1 is well-initialized, there exists a universal  
 174 constant  $c_5 > 0$  such that if

$$\kappa(\mathcal{C}) \leq c_5 \frac{\lambda_\sigma^2 r^d \nu_d}{\Lambda_\sigma \mathbb{P}(\mathcal{C}_\sigma)}, \quad (12)$$

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

176 **Remark 1.** We note that the restriction on  $\kappa(\mathcal{C})$  imposed by (12) results in a symmetric set difference  
 177  $\Delta(\mathcal{C}_\sigma[X], \widehat{C})$  on the order of  $r^d$ . In plain terms, we are able to recover a density cluster  $\mathcal{C}$  in the sense  
 178 of (5) only when we can guarantee a very small fraction of points will be misclassified. This strong  
 179 condition is the price we pay in order to obtain the uniform bound of (11).

180 **Remark 2.** While taking the radius of the neighborhood graph  $r \rightarrow 0$  as  $n \rightarrow \infty$  (thereby ensuring  
 181  $G_{n,r}$  is sparse) is computationally attractive, the presence of a factor of  $\log^2(1/r)/r$  in  $\kappa(\mathcal{C})$  unfor-  
 182 tunately prevents us from making claims about the behavior of PPR in this regime. Although the  
 183 restriction to a kernel function fixed in  $n$  is standard for spectral clustering theory [Schiebinger et al.,  
 184 2015, von Luxburg et al., 2008], it is an interesting question whether PPR exhibits some degeneracy  
 185 over  $r$ -neighborhood graphs as  $r \rightarrow 0$ , or if this is merely looseness in our upper bounds.

186 **Approximate PPR vector.** In practice, exactly solving (1) may be too computationally expensive.  
 187 To address this limitation, Andersen et al. [2006] introduced the  $\epsilon$ -approximate PPR vector (aPPR),  
 188 which we will denote by  $p^{(\epsilon)}$ . We refer the curious reader to Andersen et al. [2006] for a formal  
 189 algorithmic definition of the aPPR vector, and limit ourselves to highlighting a few salient points: the  
 190 aPPR vector can be computed in order  $\mathcal{O}(1/(\epsilon\alpha))$  time, while satisfying the following uniform error  
 191 bound:

$$\text{for all } u \in V, \quad p(u) - \epsilon D_{uu} \leq p^{(\epsilon)}(u) \leq p(u). \quad (13)$$

192 Application of (13) within the proofs of Theorems 1 and 2 leads to analogous results which hold for  
 193  $p^{(\epsilon)}$ . We formally state and prove this in the supplement.

### 194 3 Analysis

195 The primary technical contribution of our work is showing that the geometric conditions (A1)–(A5)  
 196 translate to meaningful bounds on the normalized cut and mixing time of  $\mathcal{C}_\sigma[X]$  in  $G_{n,r}$ . In doing  
 197 so, we elaborate on how some of the geometric conditions introduced in Section 2 contribute to the  
 198 difficulty of the clustering problem.

199 **Normalized cut.** We start with a finite sample upper bound on the normalized cut (3) of  $\mathcal{C}_\sigma[X]$ .  
 200 For simplicity, we write  $\Phi_{n,r}(\mathcal{C}_\sigma[X]) := \Phi(\mathcal{C}_\sigma[X]; G_{n,r})$ .

201 **Theorem 3.** Fix  $\lambda > 0$ , and assume  $\mathcal{C} \in \mathbb{C}_f(\lambda)$  satisfies Assumptions (A1)–(A3), (A5) for some  
 202  $r, \sigma, \lambda_\sigma, c_0, \gamma > 0$  (no bound on maximum density is needed). Then for any  $0 < \delta < 1$ ,  $\epsilon > 0$ , if

$$n \geq \frac{(2 + \epsilon)^2 \log(3/\delta)}{\epsilon^2} \left( \frac{25}{6\lambda_\sigma^2 \nu(\mathcal{C}_\sigma) \nu_d r^d} \right)^2, \quad (14)$$

203 then

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[X])}{r} \leq c_1 \frac{d}{\sigma} \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - c_0 \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon, \quad (15)$$

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

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

**Mixing time.** For  $S \subseteq V$ , denote by  $G[S] = (S, E_S)$  the subgraph induced by  $S$  (where the edges  
 are  $E_S = E \cap (S \times S)$ ). Let  $W_S$  be the (lazy) random walk matrix over  $G[S]$ , and write

$$q_v^{(t)}(u) = e_v W_S^t e_u$$

211 for the  $t$ -step transition probability of the lazy random walk over  $G[S]$  originating at  $v \in V$ . Also  
 212 write  $\pi = (\pi(u))_{u \in S}$  for the stationary distribution of this random walk. (As  $W_S$  is the transition  
 213 matrix of a lazy random walk, it is well-known that a unique stationary distribution exists and is given  
 214 by  $\pi(u) = (D_S)_{uu} / \text{vol}(S; G[S])$ , where we write  $D_S$  for the degree matrix of  $G[S]$ .) We define the  
 215 mixing time of  $G[S]$  as

$$\tau_\infty(G[S]) = \min \left\{ t : \frac{\pi(u) - q_v^{(t)}(u)}{\pi(u)} \leq \frac{1}{4}, \text{ for } u, v \in V \right\}. \quad (16)$$

216 Next, we give an asymptotic (in the number of vertices  $n$ ) upper bound on  $\tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]])$ .

217 **Theorem 4.** Fix  $\lambda > 0$ , and let  $\mathcal{C} \in \mathbb{C}_f(\lambda)$  satisfy Assumptions (A1) and (A4) for some  
 218  $\sigma, \lambda_\sigma, \Lambda_\sigma, \rho, L > 0$ . Then, for any  $0 < r < \sigma/2\sqrt{d}$ , with probability 1

$$\limsup_{n \rightarrow \infty} \tau_\infty(G_{n,r}[\mathcal{C}_\sigma[X]]) \leq c_2 \frac{\Lambda_\sigma^4 d^3 \rho^2 L^2}{\lambda_\sigma^4 r^2} \log^2 \left( \frac{1}{r} \right) + c_3 \quad (17)$$

219 for  $c_2, c_3 > 0$  universal constants.

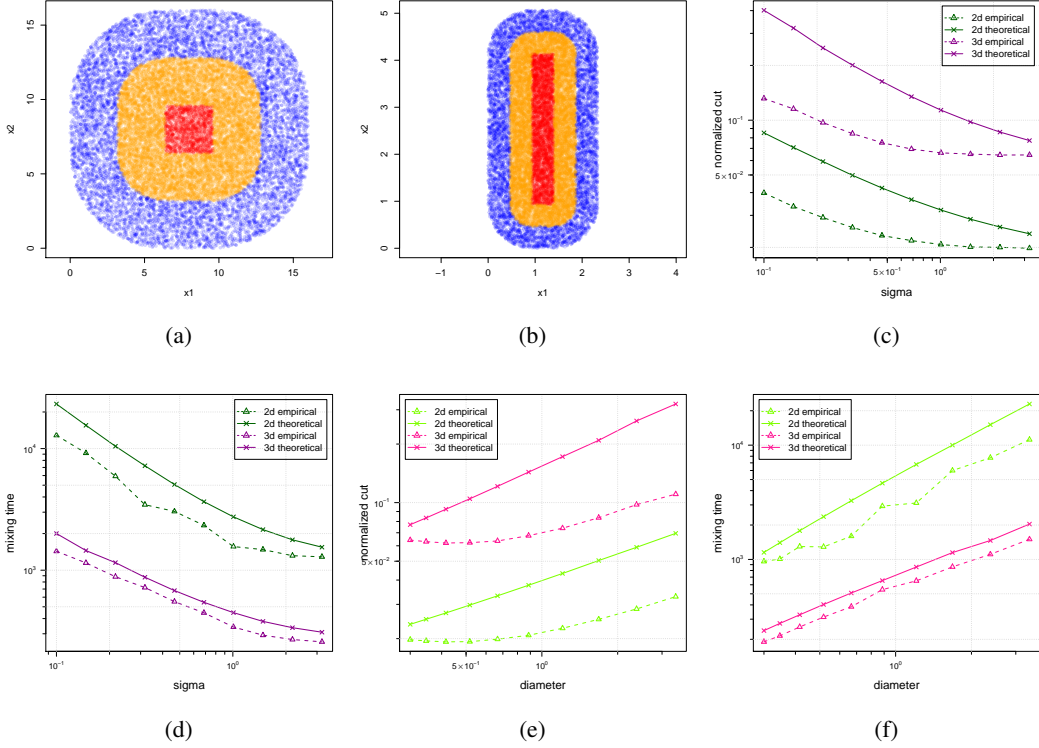


Figure 1: Samples, empirical results, and theoretical bounds for mixing time and normalized cut as diameter and thickness are varied. In (a) and (b), points in  $\mathcal{C}$  are colored in red; points in  $\mathcal{C}_\sigma \setminus \mathcal{C}$  are colored in yellow; and remaining points in blue.

To the best of our knowledge, Theorem 4 is the first bound, albeit asymptotic, on the mixing time of random walks over neighborhood graphs which is independent of  $n$ , the number of vertices.

*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  $\sigma$  and  $\rho$ , 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  $(\rho/\sigma)^d$  in (17), achieved, for instance, when  $\mathcal{C}_\sigma$  is a dumbbell-like set consisting of two balls of diameter  $\rho$  linked by a cylinder of radius  $\sigma$ . [Abbasi-Yadkori et al., 2017, Abbasi-Yadkori, 2016] develop theory regarding bi-Lipschitz 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 normalized cut and mixing time of  $\mathcal{C}_\sigma[X]$ , 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 compares these theoretical bounds with the empirical quantities (3) and (16), as we vary the diameter  $\rho$  and thickness  $\sigma$  of a cluster  $\mathcal{C}$ . Panels (a) and (b) show the resulting empirical clusters for two different values of  $\rho$  and  $\sigma$ .

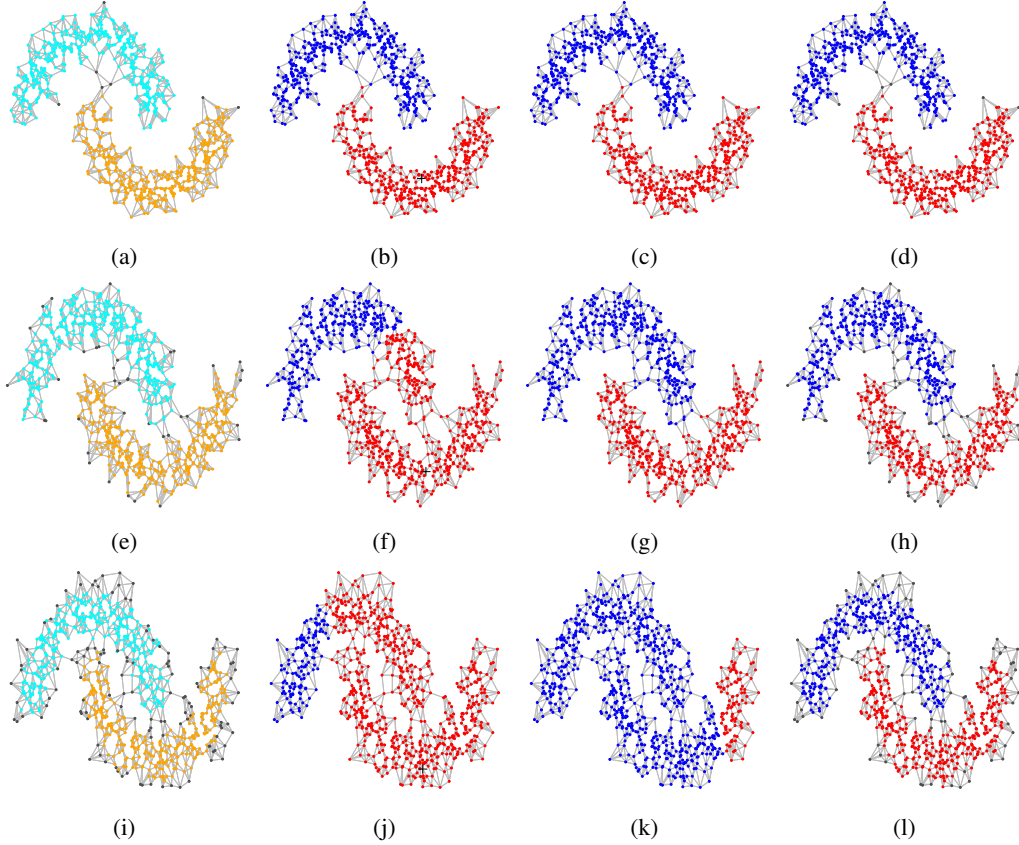


Figure 2: True density (column 1), PPR (column 2), normalized cut (column 3) and estimated density (column 4) clusters for 3 different simulated data sets. Seed node for PPR denoted by a black cross.

Panels (d) and (f) show our theoretical bounds on mixing time tracking closely with empirical mixing time, in both 2 and 3 dimensions.<sup>2</sup> This provides empirical evidence that the upper bound on mixing time given by Theorem 4 has the right dependency on both expansion parameter  $\sigma$  and diameter  $\rho$ . 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  $\sigma$  and  $\rho$  are approximately equal. As the ratio  $\rho/\sigma$  grows, we see the slopes of the empirical curves become more similar to those predicted by theory.

**Empirical behavior of PPR** 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 the well known “two moons” dataset (with added 2d Gaussian noise), considered a prototypical success story for spectral clustering algorithms. 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 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 shows the degrading ability of PPR to recover density clusters as the two moons become less well-separated. Of particular interest is the fact that PPR fails to recover one of the moons even when normalized cut still succeeds in doing so, supporting our claim from Remark 3. Additionally, we note that a density clustering algorithm recovers a moon even when both PPR and normalized

<sup>2</sup>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.



258 cut fail, lending empirical weight to our overall message that PPR recovers only geometrically  
259 well-conditioned density clusters.

## 260 **5 Discussion**

261 For given data, there are an almost limitless number of ways to define what the “right” clustering is.  
262 We have considered one such notion – density upper level sets – and have detailed a set of natural  
263 geometric criteria which, when appropriately satisfied, translate to provable bounds on estimation of  
264 the cluster by PPR. We do not, however, provide a theoretical lower bound showing that our geometric  
265 conditions are required for successful recovery on an upper level set. Although we investigate the  
266 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 l-spectral clustering and sparse pca. In *Advances in Neural Information Processing Systems*, pages 847–855, 2010.
- Aleksandr P. Korostelev and Alexandre B. Tsybakov. *Minimax theory of image reconstruction*. Springer, 1993.
- 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.
- Wolfgang Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.
- Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.
- Geoffrey Schiebinger, Martin J. Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *Ann. Statist.*, 43(2):819–846, 04 2015.

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

316 Aarti Singh, Clayton Scott, and Robert Nowak. Adaptive hausdorff estimation of density level sets.  
317 *Ann. Statist.*, 37(5B):2760–2782, 10 2009.

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

320 Daniel A Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its  
321 application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26,  
322 2013.

323 Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and  
324 solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and*  
325 *Applications*, 35(3):835–885, 2014.

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

327 Alexandre B Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*,  
328 25(3):948–969, 1997.

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

331 Xiao-Ming Wu, Zhenguo Li, Anthony M. So, John Wright, and Shih fu Chang. Learning with partially  
332 absorbing random walks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors,  
333 *Advances in Neural Information Processing Systems 25*, pages 3077–3085. Curran Associates, Inc.,  
334 2012.

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