Local Spectral Clustering of Density Upper Level Sets

Anonymous Authors[1]

Abstract

1. Introduction

Let $\mathbf{X} := (x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Our statistical learning task is clustering: splitting data into groups which satisfy some notion of within-group similarity and between-group difference.

In particular, spectral clustering methods are a family of powerful non-parametric clustering algorithms. Given a symmetric adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with (i, j)th entry representing the similarity between data points x_i and x_j, we form the random walk transition probability matrix $\mathbf{W}$, and corresponding graph Laplacian matrix $\mathbf{L}$[1]:

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

where the degree matrix $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} := \sum_j \mathbf{A}_{ij}$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

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

A natural model to consider when analyzing point cloud data such as $\mathbf{X}$ is the following:

$$x_i \sim \mathbb{P}, \quad \text{independently, for } i = 1, \ldots, n, \tag{2}$$

with f the density function of $\mathbb{P}$ with respect to the uniform measure over $\mathbb{R}^d$. In this case, we are interested in understanding what the output of a clustering algorithm on this finite sample reveals about the unknown density f. For $\lambda > 0$ and the upper level set $\{x : f(x) \geq \lambda\}$, it is intuitive (Hartigan, 1981; Chaudhuri & Dasgupta, 2010) to define clusters as the connected components $\mathbb{C}_f(\lambda)$ of the upper level set; we call these connected regions of high density *density clusters*, and study the ability of spectral methods to identify such clusters.

Graph connectivity criteria. A somewhat more standard mode of understanding spectral clustering methods is to view them as approximating some graph connectivity criteria.

For $\mathbf{W}$ as before, let the graph $G = (V, E)$, with vertices $V = \{v_1, \ldots, v_n\}$ corresponding to the n rows of $\mathbf{A}$, and (possibly weighted) edges $E = \{(v_i, v_j, \mathbf{A}_{ij}) : 1 \leq i < j \leq n, \mathbf{A}_{ij} > 0\}$ (Since $\mathbf{A}$ is symmetric, G is an undirected graph; also, by convention, we preclude self-loops, hence $i \neq j$). There are many (Yang & Leskovec, 2015; Fortunato, 2010) graph-theoretic measures which assess the cluster quality of a subset $S \subseteq V$ (or more generally the quality of a partition $S_1 \cup \ldots \cup S_m = V$, for $m \geq 2$.)

Arguably a natural way to assess cluster quality is via a pair of criteria capturing the *external* and *internal connectivity* of S, respectively. As the names suggest, external connectivity should relate to the number of edges between S and its complement G/S (hereafter denoted S^c), while internal connectivity in turn measures the number of edges between subsets within S. The graph clustering task then becomes to find a subset S (or, for global algorithms, a partition $S_1 \cup \ldots \cup S_m = V$), which has both small external and large internal connectivity.

We will assess the external connectivity of a subset $S \subseteq V$ through its normalized cut. The cut of S is

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

– where $S^c = V/S$ is the complement of S in V – and the volume of S is

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

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

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

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

Then, the *normalized cut* of S is given by

$$\Phi(S;G) := \frac{\text{cut}(S;G)}{\min\{\text{vol}(S;G), \text{vol}(S^c;G)\}} \tag{3}$$

Intuitively, a set with low *normalized cut* has many more edges which do not cross the cut than edges which do cross the cut.

Given $S \subseteq V$, the subgraph induced by S is given by $G[S] = (S, E_S)$, where $(u,v) \in E_S$ if both u and v are in S and $(u,v) \in E_S$. Letting $|S| = m$, $\mathbf{A}_S$ then denotes the $m \times m$ adjacency matrix representation of $G[S]$; similarly, $\mathbf{D}_S$ is the diagonal degree matrix with entries $(\mathbf{D}_S)_{ii} = \sum_{j:v_j \in S} \mathbf{A}_{ij}$, and $\mathbf{W} = \mathbf{D}_S^{-1}\mathbf{A}_S$ is the corresponding random walk matrix (again, only over $G[S]$.)

Our internal connectivity parameter $\Psi(S)$ will capture the time it takes for the random walk governed by $\mathbf{W}_S$ to mix (that is, approach a stationary distribution) uniformly over S. Denoting the stationary distribution $\widetilde{\boldsymbol{\pi}} = (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_m)$, the *relative pointwise mixing time* $\tau_\infty(S; G[S])$ of the random walk over $G[S]$ defined by $\mathbf{W}_S$ is the smallest integer $t_0 > 0$ such that for all $v = v_i, v' = v_j \in S$:

$$\left| \frac{e_v \mathbf{W}_S^t - \widetilde{\pi}_j}{\widetilde{\pi}_j} \right| \leq \frac{1}{4}$$

for all $t > t_0$ [2] (Of course, given the definition of $\mathbf{W}_S$ it is well known that the stationary distribution $\boldsymbol{\pi}_S$ will be defined by $\widetilde{\pi}_i = (\mathbf{D}_S)_{ii}/\text{vol}(S; G[S])$.

Intuitively, the smaller the pointwise mixing time, the more connected every pair of points v and v' are in the graph $G[S]$. Therefore, the internal connectivity parameter $\Psi(S;G)$ is simply one over the mixing time:

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

If S has normalized cut no greater than Φ, and inverse mixing time no less than Ψ, we will refer to it as a (Φ, Ψ)-cluster. Both local (Zhu et al., 2013) and global (Kannan et al., 2004) spectral algorithms have been shown to output clusters (or partitions) which provably satisfy approximations to the optimal (Φ, Ψ)-cluster (or partition), where the optimization is carried out over the graph G. [3]

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

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

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

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

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

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

To compute a cluster $\widehat{C} \subset V$ using the PPR vector, we will take sweep cuts of $\mathbf{p}(v, \alpha; G)$. Denote the entries of $\mathbf{p}(v, \alpha; G)$ by $\mathbf{p}(v, \alpha; G) = (p_1, \ldots, p_n)$, and let the *stationary distribution* $\boldsymbol{\pi}$ of the random walk defined by $\mathbf{W}$ be given by

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n), \quad \pi_j := \frac{\mathbf{D}_{jj}}{\text{vol}(V; G)}.$$

Then, for a number $\beta \geq 0$, the sweep cut S_β is

$$S_\beta = \{u_j \in V : p_j > \beta \pi_j\}. \tag{6}$$

One such sweep cut will be our chosen cluster $\widehat{C}$. (We delay formal introduction of the local clustering algorithm we analyze until Section 2, after we have given a method for forming a graph over the data $\mathbf{X}$.)

Large sample behavior. Let (r_n) be a sequence of positive numbers. Given a sequence of kernel functions $k_n : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ of the form $k_n(x, x') = k(\|x - x'\| / r_n)$ for k a non-increasing function, and data $\mathbf{X} = \{x_1, \ldots, x_n\}$ sampled from $\mathbb{P}$ as before, form the (weighted, complete) similarity graph $G_n = (\mathbf{X}, E_n)$ with

$E_n = \{k(x_i, x_j) : 1 \leq i < j \leq n\}$. (Here, $\|\cdot\|$ is used to denote Euclidean norm).

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

1.1. Summary of results

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

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

and the associated *neighborhood graph*

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

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

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

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

occur with probability tending to 1.

Our results can now be summarized by the following two points:

1. Under a natural set of geometric conditions[4], the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$ can be bounded. Theorems 1 and Theorem 2 provide an upper and lower bound, respectively
2. We show these bounds on the graph connectivity criteria have algorithmic consequences personalized PageRank. An immediate consequence of Theorems 1 and Theorem 2, along with the previous work of (Zhu et al., 2013), is to yield an upper bound on the normalized cut of the set $\widehat{\mathcal{C}}_n$ output by Algorithm 1, as well as upper bounding the symmetric set difference between $\widehat{\mathcal{C}}_n$ and $\mathcal{C}[\mathbf{X}]$. Furthermore, in Section 4 we show that a careful analysis of the form typical to local clustering algorithms yields Theorem 4, which states that Algorithm 1, properly initialized, performs consistent density cluster estimation in the sense of (9).

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

Organization. In Section 4, we provide some example density functions, to clarify the relevance of our results. In Section 6, we show empirical performance of the PPR algorithm, which demonstrates that violations of the geometric conditions we set out in Section **??** manifestly impact density cluster recovery (i.e. the conditions are not superfluous), before concluding in Section 7. First, however, we summarize some related work.

1.2. Related Work

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

Global spectral clustering methods were first developed in the context of graph partitioning (Fiedler, 1973; Donath & Hoffman, 1973) and their performance is well-understood in this context (see, for instance, (Tolliver & Miller, 2006; von Luxburg, 2007)). In a similar vein, several recent works (McSherry, 2001; Lei & Rinaldo, 2015; Rohe et al., 2011; Abbe, 2018; Chaudhuri et al., 2012; Balakrishnan et al., 2011) have studied the efficacy of spectral methods in successfully recovering the community structure in various variants of the stochastic block model.

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

Perhaps most similar to our results are (Vempala & Wang, 2004; Shi et al., 2009; Schiebinger et al., 2015), which study the consistency of spectral algorithms in recovering the latent labels in certain parametric and non-parametric mixture models. These results focus on global rather than local algorithms, and as such impose global rather than local

conditions on the nature of the density. Moreover, they do not in general ensure recovery of density clusters which is a focus of our work.

2. Background and Assumptions.

We begin by introducing an defining well-conditioned density clusters before turning to formally define the PPR algorithm under consideration, Algorithm 1, and discussing the choice of tuning-parameters.

2.1. Well-conditioned density clusters.

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

Let $B(x,r) = \{y \in \mathbb{R}^d : \|y-x\| \leq r\}$ be a closed ball of radius r around the point x. Given a set $\mathcal{A} \subset \mathbb{R}^d$, and a number $\sigma > 0$, define the set $\mathcal{A}_\sigma = \mathcal{A} + B(0,\sigma) = \{y \in \mathbb{R}^d : \inf_{x\in\mathcal{A}} \|y-x\| \leq \sigma\}$. The assumptions we require are as follows:

(A1) *Minimum density:* For numbers $\sigma, \lambda_\sigma > 0$, a set $\mathcal{A} \subset \mathbb{R}^d$ is said to be (σ,λ_σ)-*regular* if

$$\inf_{x\in\mathcal{A}_\sigma} f(x) = \lambda_\sigma \tag{10}$$

(A2) *Low noise density:* For numbers $\gamma, \sigma > 0$, and a set $\mathcal{A} \subset \mathbb{R}^d$, the density function f is said to be (γ,σ)-*low noise around* $\mathcal{A}$ if there exists a constant $c_1 > 0$ such that for all $x \in \mathbb{R}^d$ with $0 < \rho(x,\mathcal{A}_\sigma) \leq \sigma$,

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

where $\rho(x,\mathcal{A}_\sigma) = \min_{x_0\in\mathcal{A}_\sigma} \|x-x_0\|$.

(A3) *Cluster separation:* For $\lambda, \sigma > 0$, $\mathcal{C} \in \mathbb{C}_f(\lambda)$ is said to be σ-*well separated* if for all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$,

$$\rho(\mathcal{C},\mathcal{C}') > \sigma.$$

where $\rho(\mathcal{C},\mathcal{C}') = \min_{x\in\mathcal{C}} \rho(x,\mathcal{C}')$,

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

Definition 2. *For $\lambda > 0$ we say $\mathcal{C} \in \mathbb{C}_f(\lambda)$ is a $(\lambda_\sigma,\gamma,\sigma)$-well conditioned cluster if it satisfies Assumptions (A1)-(A3), for some $\lambda_\sigma, \gamma, \sigma > 0$.*

We note that σ plays dual roles here, both in effect precluding arbitrarily narrow and long clusters in (A1) and arbitrarily flat densities in (A2).

2.2. Algorithm under consideration.

Algorithm 1 will be the simple procedure we analyze. It will take as input the data $\mathbf{X}$ along with user-specified parameters r, α, vol_0, and $v \in \mathbf{X}$, and perform the following steps:

Algorithm 1 PPR on a neighborhood graph

Input: data $\mathbf{X}$, radius r, teleportation parameter $\alpha \in [0,1]$, seed node $v \in \mathbf{X}$, target volume vol_0.
Output: $\widehat{C} \subset V$.

1: Form the neighborhood graph $G_{n,r}$ as given in (8)
2: Compute PPR vector $\mathbf{p}(v,\alpha; G_{n,r})$ as defined by (5).
3: For $\beta \in [\frac{1}{8}, \frac{1}{2}]$ compute sweep cuts S_β as defined by (6).
4: Return

$$\widehat{C} = \underset{\beta\in[\frac{1}{8},\frac{1}{2}]}{\arg\min}\, \Phi(S_\beta; G_{n,r})$$

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.

Define a well-initialized PPR algorithm

Make clear via assumption that throughout we require the algorithm to be well-initialized.

For notational simplicity, hereafter for $S \subseteq \mathbf{X}$ we will refer to $\Phi(S; G_{n,r})$ as $\Phi_{n,r}(S)$, and likewise with $\Psi(S; G_{n,r})$ and $\Psi_{n,r}(S)$.

3. Local Clustering on Density Level Sets

In this section we provide bounds for the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$. As a result we can lower bound $\Phi(\widehat{C}; G_{n,r})$ for $\widehat{C}$ the output of a well-initialized PPR algorithm, and lower bound the symmetric set difference between $\widehat{C}$ and $\mathcal{C}[\mathbf{X}]$.

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

- ***Additive error bound.*** *Fix $\epsilon > 0$. Then, for*

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

we have

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

- ***Multiplicative error bound.***

where $c_0 = 2^{2d+1} d$ and $c_\sigma = c_0/\sigma$.

The proof of Theorem 1, along with all other theorems, can be found in the supplementary document. A few remarks are in order.

Remark 1. Note that the bound of (12) depends exponentially on d; precisely, on $C_0 = d2^{2d+1}$. It is possible to improve this dependency to the order of $(1 + \frac{r}{\sigma})^{2d}$. However, in this setting, we think of r as being a constant radius (rather than $r = r_n \to 0$ as $n \to \infty$, and therefore even this improvement still retains an exponential dependency on the dimension d.

Remark 2. Aside from the looseness implied by Remark 1, the error bound of (12) is almost tight. Specifically, choosing

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

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

$$\frac{\Phi_{n,r}(\mathcal{A}_\sigma[\mathbf{X}])}{r} \geq c_1 \frac{(\lambda - \frac{r^{\epsilon+1}}{\epsilon+1})}{\lambda} - \epsilon \tag{13}$$

for some constant c_1 which depends only on dimension. (Note that a factor of $1/\sigma$ is not (12) replicated in this lower bound.) Provide justification in supplement, and cite it.

We now provide an upper bound for $\Psi_{n,r}(\mathcal{C}[\mathbf{X}])$.

Theorem 2. *For c_d a constant which may depend only on the dimension d,*

$$\Psi_{n,r}(\mathcal{C}[\mathbf{X}]) \leq c_d \frac{\sigma^d}{D^d} \frac{\lambda_{\min}^7}{\lambda_{\max}^7} \frac{1}{\log(\lambda_{\max}^2/(\lambda_{\min} r^d \nu_d))}$$

4. Implications, extensions, and discussion

4.1. Examples

4.2. Experiments

References

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

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

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

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

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

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

Ery, A.-C., Pelletier, B., and Pudlo, P. The normalized graph cut and cheeger constant: from discrete to continuous. *Adv. in Appl. Probab.*, 44(4):907–937, 12 2012. doi: 10.1239/aap/1354716583. URL https://doi.org/10.1239/aap/1354716583.

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. doi: https://doi.org/10.1016/j.physrep.2009.11.002. URL http://www.sciencedirect.com/science/article/pii/S0370157309002841.

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.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004. ISSN 0004-5411. doi: 10.1145/990308.990313. URL http://doi.acm.org/10.1145/990308.990313.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Trillos, N. G., Slepčev, D., Von Brecht, J., Laurent, T., and Bresson, X. Consistency of cheeger and ratio graph cuts. *J. Mach. Learn. Res.*, 17(1):6268–6313, January 2016. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2946645.3053463.

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

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

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

Yang, J. and Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, Jan 2015. ISSN 0219-3116. doi: 10.1007/s10115-013-0693-z. URL https://doi.org/10.1007/s10115-013-0693-z.

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