Spectral Methods for Testing Clusters in Graphs

Sandeep Silwal, Jonathan Tidor
Mentor: Jonathan Tidor

August 31, 2018

Abstract

We study the problem of testing if a graph has a cluster structure in the framework of graph property testing. Given some ϵ, we say that a graph with a degree bound d is (k, ϕ) clusterable if it can be partitioned into at most k parts such that the inner conductance of each part is at least ϕ and the outer conductance of each part is at most $c_{d,k}\phi^2$. We present an algorithm that accepts all graphs that are $(k, \phi)-$clusterable with probability at least $\frac{2}{3}$ and rejects all graphs that are ϵ-far from (k, ϕ^*)-clusterable for $\phi^* \leq c_{d,k}\mu\phi^2\epsilon^2$ where $0 < \mu < C$ for some constant C. This improves upon the work of Czumaj, Peng, and Sohler in [2] by removing a $\log n$ factor from the denominator of ϕ^*. Our technique combines the ideas present in [2, 7] in addition to a new spectral perspective. Currently this edition of the paper only has a proof for the case of $k = 2$ but the proof for all k will be given in a future edition of the paper.

1 Introduction

1.1 Background on Property Testing

In this paper, we study property testing of graphs in the bounded degree model. We are given a graph $G = (V, E)$ on n vertices where all the vertices have degree at most d. Given a graph property $\mathcal{P}$, we say that G is ϵ-far from satisfying $\mathcal{P}$ if ϵdn edges need to be added or removed from G for G to satisfy $\mathcal{P}$. A property testing algorithm for $\mathcal{P}$ is an algorithm that accepts every graph G with $\mathcal{P}$ with probability at least $\frac{2}{3}$ and rejects every graph that is ϵ-far from having $\mathcal{P}$ with probability at least $\frac{2}{3}$.

We assume that the representation of G is given as an oracle which allows us to find the ith neighbor of any vertex v if $i \leq d$. If i is larger than the degree of v, then a special symbol is returned. The complexity of a graph property testing algorithm is measured in terms of its query complexity, or the number of queries the algorithm accesses from the oracle. Therefore, the goal of property testing algorithms is to find an algorithm with an efficient query complexity. This framework of property testing of graphs was developed by Goldreich and Ron [4]. has been applied to study various properties such as bipartiteness [5] and 3-colorability [4]. See [4] and [9] for some more examples.

Our paper deals with a slightly different definition of property testing. We test for a property $\mathcal{P}$ that is parameterized by a parameter α such that $\mathcal{P}(\alpha) \subseteq \mathcal{P}(\alpha')$ if $\alpha \leq \alpha'$. We

then accept graphs with property $\mathcal{P}(\alpha)$ with probability at least $\frac{2}{3}$ and reject graphs that are ϵ-far from having property $\mathcal{P}(\alpha')$ with probability at least $\frac{2}{3}$ where $\alpha < \alpha'$.

1.2 Testing k-clusterability

We are interested in the property of k clusterability as introduced by Czumaj, Peng, and Sohler in [2]. Roughly, a graph is k-clusterable if it can be partitioned into at most k clusters where vertices in the same cluster are 'well connected' while vertices in different clusters are 'well-separated.' The quality of the clusters will be measured in terms of their inner and outer conductance, which are defined below. The idea of using conductance for graph clustering has been studied in numerous works, such as [10].

More formally, if $S \subset V$ such that $|S| \leq |V|/2$, the conductance of S is defined as $\phi_G(S) = \frac{e(S,V\setminus S)}{d|S|}$ where $e(S, V \setminus S)$ is the number of edges between S and $(V \setminus S)$. The conductance of G is defined to be the minimum conductance over all subsets $|S| \leq V/2$ and is denoted as $\phi(G)$. Now for any $S \subseteq V$, let $G[S]$ denote the induced subgraph of G on the vertex set defined by S. We will let $\phi(G[S])$ denote the conductance of this subgraph. To avoid confusion, if $S \subseteq V$, we will call $\phi_G(S)$ the *outer conductance* of S and $\phi(G[S])$ the *inner conductance*.

We say G is (k, ϕ)-clusterable if there exits a partition of V into at most k subsets C_i such that $\phi(G[C_i]) \geq \phi$ but $\phi(C_i) \leq c_{d,k}\phi^2$ for all i. where $c_{d,k}$ is a constant that depends only on d and k. Czumaj et al. have created an algorithm that accepts all (k, ϕ)-clusterable graphs with probability at least $\frac{2}{3}$ and rejects all graphs that are ϵ-far from (k, ϕ^*)-clusterable graphs where $\phi^* = c'_{d,k}\frac{\phi^2\epsilon^4}{\log n}$, where $c_{d,k}$ depends only on d, k. Our work improves upon this result by removing the $\log n$ dependency. We present an algorithm that accepts all (k, ϕ)-clusterable graphs with probability at least $\frac{2}{3}$ and rejects all graphs that are ϵ-far from (k, ϕ^*)-clusterable with probability at least $\frac{2}{3}$ where $\phi^* = c''_{d,k}\phi^2\epsilon^2$. In this writeup, we only present the case of $k = 2$ and our paper on a general k will be available on the arxiv shortly. We rely on spectral tools to remove the $\log n$ dependency. Specifically, our algorithm works by sampling a minor of a power of M, the lazy random walk matrix. We then reject the graph if all eigenvalues of this matrix are 'large' and accept otherwise. Our main result is the following.

Theorem 1.1. *Let $c_{d,k}$ be a constant depending on d and k. Then, Algorithm 2.1 accepts every (k, ϕ)-clusterable graph of maximum degree at most d with probability at least $\frac{2}{3}$ and and rejects every graph of maximum degree at most d that is ϵ-far from being (k, ϕ^*)-clusterable with probability at least $\frac{2}{3}$ if $\phi^* \leq \mu c_{d,k}\phi^2\epsilon^2$ and $0 < \mu < C$ for some constant C.*

As stated above, the current writeup will only present the case of $k = 2$ with the analysis for a general k in a forthcoming paper.

1.3 Related Work

Testing k-clusterability was inspired by property testing of expansions which has been well studied. A graph is called an α expander if if every $S \subset V$ of size at most $V/2$ has a neighborhood of size at least $\alpha|S|$. Czumaj and Sohler [3] showed that an algorithm of Goldreich and Ron [6] can distinguish between α expanders of degree bound d and graphs

which are ϵ far from having expansion at least $\Omega(\alpha^2/\log(n))$. This work was subsequently improved by Kale and Seshadhri [7] and then by Nachmias and Shapira [8] who showed that the algorithm of Goldreich and Ron can actually distinguish graphs which are α expanders between graphs which are ϵ-far from $\Omega(\alpha^2)$ expanders.

The work of Czumaj et al. for testing k-clusterability also uses property testing of distributions, such as testing the l_2 norm of a discrete distribution and testing the closeness of two discrete distributions are both used. For some work on testing the norm of a discrete distribution and testing closeness of discrete distributions, see [3] and [1] respectively.

1.4 Definitions

In this section we introduce some definitions and tools that we will be used in our analysis. Given a graph G with degree bound d, we let $\mathbf{M}$ denotes its lazy random walk matrix and $2I - 2\mathbf{M}$ denotes its Laplacian matrix $\mathbf{L}$. Let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 2$ denote the eigenvalues of L and let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ denote the corresponding orthonormal eigenvectors. Let $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_n$ denote the eigenvalues of $\mathbf{M}$ where $\nu_i = 1 - \frac{\lambda_i}{2}$ for $1 \leq i \leq n$.

For $u \in G$, we will denote $\mathbf{p}_u^t$ as the probability distribution of the endpoint of a length t random walk that starts at vertex u. That is,

$$\mathbf{p}_u^t = \mathbf{1}_u \mathbf{M}^t = \sum_{i=1}^{n} \mathbf{v}_i(u) \left(1 - \frac{\lambda_i}{2}\right)^t \mathbf{v}_i.$$

It will also be convinient to expander our definition of (k, ϕ)-clusterable. For an undirected graph G, and parameters k, ϕ_{in}, ϕ_{out}, we define G to be $(k, \phi_{in}, \phi_{out})$-clusterable if there exists a partition of V into h subsets $C_1, \cdots C_h$ such that $1 \leq h \leq k$ and for each i, $1 \leq i \leq h, \phi(G[C_]) \geq \phi_{in}$ and $\phi_G(C_i) \leq \phi_{out}$. Hence in our case, we want to accept graphs that are $(k, \phi, c_{d,k}\phi^2)$-clusterable and reject graphs that are $\epsilon-$far from $(k, \mu\phi^2\epsilon^2, c'_{d,k}\phi^4\epsilon^4)-$clusterable.

In this paper, we will denote $\|\cdot\|$ to denote the l_2 norm. We will also need the following result from [11] which roughly states that eigenvalues are stable under a good approximation.

Theorem 1.2. *Let $H = A + P$ and suppose H has eigenvalues $\mu_1 \geq \cdots \geq \mu_n$ and A has eigenvalues $\nu_1 \geq \cdots \geq \nu_n$. Furthermore, suppose $\|P\|_2 \leq \epsilon$. Then, $|\mu_i - \nu_i| \leq \epsilon$ for all $1 \leq i \leq n$.*

Our algorithm relies on estimating dot products and norms of various distributions. Therefore, we need the following results about distribution property testing.

Theorem 1.3. *[1, Theorem 1.2] Let $\delta, \xi > 0$ and let $\mathbf{p}, \mathbf{q}$ be two distributions over a set n with $b \geq \max(\|\mathbf{p}\|^2, \|\mathbf{q}\|^2)$. Let $r > \Omega\left(\frac{\sqrt{b}}{\xi} \log \frac{1}{\eta}\right)$. Then there exists an algorithm denoted by l_2-**Inner-Product-Estimator** that takes as input r samples from each distribution $\mathbf{p}, \mathbf{q}$, and returns an estimate of $\langle p, q \rangle$ that is accurate to within $O(\xi)$ with probability at least $1-\eta$.*

Theorem 1.4. *[2, Lemma 3.2] Let $G = (V, E)$ with $|V| = n$. Let $v \in V, \sigma > 0$, and $r \geq 16\sqrt{n}$. There exists an algorithm denoted as l_2-**Norm-Tester** that takes as input r samples from $\mathbf{p}_v^t$ and accepts the distribution of $\|\mathbf{p}_v^t\|^2 \leq \frac{\sigma}{4}$ and rejects the distribution if $\|p_v^t\| > \sigma$, with probability at least $1 - \frac{16\sqrt{n}}{r}$.*

2 The Algorithm

In this section, we describe our algorithm referenced in Theorem 1.1. As stated section 1.2 , our algorithm performs lazy random walks and uses the distribution testing results 1.3 and 1.4 to approximate a minor of a power of M. For all v in V, we define $\mathbf{q}_v^t = \mathbf{p}_v^t - \frac{1}{n}\mathbf{1}$. We also define the minor $A_{u,v}$ of $M^{2t} - \frac{1}{n}\mathbf{J}$ as

$$A_{u,v} = \begin{bmatrix} \|\mathbf{q}_u\|^2 & \langle \mathbf{q}_u, \mathbf{q}_v \rangle \\ \langle \mathbf{q}_v, \mathbf{q}_u \rangle & \|\mathbf{q}_v\|^2 \end{bmatrix}.$$

Our algorithm is the following:

Algorithm 2.1: Cluster-Test$(G, S, N, t, \sigma, \mu, \xi)$

for S *rounds* **do**
  Pick a pair of vertices u and v uniformly at random from G.
  Run N random walks of length t.
  Using the results of 3 as an input, test if either $\|\mathbf{p}_u^t\|^2 \geq \sigma$ or $\|\mathbf{p}_v^t\|^2 \geq \sigma$ through the l_2 − **Norm-Tester** as defined in 1.4. If so, abort and reject.
  Using the results of 3 as an input and the l_2**-Inner-Product-Estimator**, approximate each entry of $A_{u,v}$ within ξ. Call the approximation $\tilde{A}_{u,v}$.
  Abort and reject if both the eigenvalues of $\tilde{A}_{u,v}$ are larger than $\frac{3}{n^{1+\mu}}$.
Accept.

3 Proof of Theorem 1.1.

We will now prove our main result, theorem 1.1. In the **Cluster-Test**, we set $\theta < \frac{3-2\sqrt{2}}{48}$, $S = \frac{128 \cdot 1152^2}{\epsilon^4 \theta^2}$, $0 < \mu < C$ where $C = \min(1, \frac{1}{2(512 c_{4.5} c_{5.1}^2 - 1)})$, $\gamma = \eta = \frac{1}{8S}$, $r = \frac{16\sqrt{n}}{\eta}$, $N = O(n^{1/2+\mu})$, $t = \frac{64 c_{5.1}^2 \log(n)}{\phi^2}$, $\sigma = \frac{4}{\gamma n}$.

3.1 Completeness: Accepting $(k, \phi)-$clusterable Graphs.

We will show that the **Cluster-Test** algorithm with the parameters defined above accepts G with probability greater than $\frac{2}{3}$ if G is $(2, \phi, c_d\phi^2)-$clusterable. We will do this by showing that if G is 2-clusterable, then the vectors $\mathbf{q}_v^t$ are essentially 'one dimensional.' Hence, the matrices $A_{u,v}$ should be close to rank 1 which by Theorem 1.2, at least one eigenvalue of $A_{u,v}$ should be sufficiently small. First we need a result from [2] that states that there is a gap between the kth and the $k+1$th eigenvalue of M if G is $(k, \phi_{in}, \phi_{out})$-clusterable.

Lemma 1. *[2, Lemma 5.2] If G is a weighted d-regular graph and $(k, \phi_{in}, \phi_{out})$-clusterable then there exists h, $1 \leq h \leq k$, and a constant $c_{5.1}$ such that $\lambda_i \leq 2\phi_{out}$ for any $i \leq h$ and $\lambda_i \geq \frac{\phi_{in}^2}{c_{5.1}^2 h^4}$ for any $i \geq h+1$.*

We will also use the following result from [2] which states that $(k, \phi_{in}, \phi_{out})$-clusterable are accepted with sufficiently high probability in step 4 of Algorithm 2.1.

Lemma 2. *[2, Lemma 4.3]. Let $0 < \gamma < 1$. If G is $(k, \phi_{in}, \phi_{out})$-clusterable, then there $V' \subseteq V$ with $|V'| \geq (1-\gamma)|V|$ such that for any $u \in V'$ and any $t > \frac{c_{4.3}k^4 \log(n)}{\phi_{in}^2}$, for some universal constant $c_{4.3}$, the following holds:*

$$\left\|\mathbf{p}_u^t\right\|^2 \leq \frac{2k}{\gamma n}.$$

The next two lemmas tell us that for $(k, \phi_{in}, \phi_{out})$-clusterable graphs, $\mathbf{q}_u = \mathbf{p}_u - \frac{1}{n}\mathbf{1}$ is well approximated by its projection onto $\mathbf{v}_2$ for all vertices u in G.

Lemma 3. *Let G be $(2, \phi_{in}, \phi_{out})$-clusterable. For a vertex u, let $\mathbf{p}_u'^t$ be the projection of $\mathbf{p}_v$ onto the space spanned by $\mathbf{v}_3, \ldots, \mathbf{v}_n$. Then $\|\mathbf{p}_u^t - \mathbf{p}_u'^t\|^2 \leq \left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^{2t}$.*

Proof. Write p_u^t in the eigenbasis of L. Then,

$$\begin{aligned}
\left\|\mathbf{p}_u^t - \mathbf{p}_u'^t\right\|^2 &= \sum_{i=3}^{n} \left(1 - \frac{\lambda_i}{2}\right)^{2t} \mathbf{v}_i(u)^2 \\
&\leq \left(1 - \frac{\lambda_3}{2}\right)^{2t} \sum_{i=3}^{n} \mathbf{v}_i(u)^2 \\
&\leq \left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^{2t}.
\end{aligned}$$

□

We can rewrite the conclusion of lemma 3 in a slightly more useful statement.

Lemma 4. *Let u and v be two vertices. Then we can find $\mathbf{e}_u$ and $\mathbf{e}_v$ such that both $\mathbf{q}_u^t + \mathbf{e}_u$ and $\mathbf{q}_v^t + \mathbf{e}_v$ lie on a line through the origin and $\max(\|\mathbf{e}_u\|^2, \|\mathbf{e}_v\|^2) < \left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^{2t}$.*

Lemma 5. *Let G be $(2, \phi_{in}, \phi_{out})$-clusterable and let u, v be any two vertices in G. Then $A_{u,v}$ has an eigenvalue less than $\sqrt{2}\left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^t$.*

Proof. We know that $A_{u,v}$ is a gram matrix and hence positive semi-definite. Therefore, we can write $A_{u,v} = \kappa_1^2 w_1 w_1^T + \kappa_2^2 w_2 w_2^T$ where $\langle w_1, w_2 \rangle = 0$. Now let M_1 be a $n \times 2$ matrix with columns $\mathbf{q}_u$ and $\mathbf{q}_v$ and define $M_2 = \kappa_1 w_1 w_1^T + \kappa_2 w_2 w_2^T$. Denote the columns of M_2 as $\mathbf{q}'_u, \mathbf{q}'_v$. By the orthogonality of w_1 and w_2, we have that $M_1^T M_1 = M_2^T M_2$. Hence, $U = M_1 M_2^{-1}$ is an orthogonal matrix and the image of $\mathbf{q}'_u$ is $\mathbf{q}_u$ and similarly, the image of $\mathbf{q}'_v$ is $\mathbf{q}_v$.

Now by Lemma 4, we know we can find $\mathbf{e}_u$ and $\mathbf{e}_v$ such that $\|\mathbf{e}_u\|, \|\mathbf{e}_v\| < \left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^t$ such that M_1 plus the matrix with columns $\mathbf{e}_u, \mathbf{e}_v$ is rank 1. Since U is orthogonal, we can find $\mathbf{e}'_u, \mathbf{e}'_v$ such that $\|\mathbf{e}'_u\|, \|\mathbf{e}'_v\| < \left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^t$ such that M_2 plus the matrix with columns $\mathbf{e}'_u, \mathbf{e}'_v$ is rank 1. By Weyl's inequality (Theorem 1.2), this means that one of κ_1, κ_2 must be smaller than $\sqrt{2}\left(1 - \frac{\phi_{in}^2}{32c_{5.1}^2}\right)^t$. □

Combining the above lemmas, we get the following lemma.

Lemma 6. *For the parameters specified in section 3, Algorithm 2.1 accepts $(2, \phi, c_d\phi^2)$-clusterable graphs with probability greater than $\frac{2}{3}$.*

Proof. From Lemma 2, we know that in step 4 of the algorithm, we accept with probability $(1-\gamma)^2$. After this step, we know that by Lemma 5 that $A_{u,v}$ has an eigenvalue that is at most $\sqrt{2}\left(1-\frac{\phi_{in}^2}{32c_{5.1}^2}\right)^t \leq \sqrt{2}n^{-2}$. When we estimate $A_{u,v}$ with $\tilde{A}_{u,v}$ in step 5 of Algorithm 2.1, the eigenvalues of A shift by at most $\frac{3}{n^{1+\mu}}$ by Weyl's inequality [11]. Then by the union bound, the probability that we reject is at most

$$S(1-(1-\gamma)^2+\gamma^2\eta) < \frac{1}{4}+\frac{1}{64} < \frac{1}{3}.$$

□

3.2 Soundness: Rejecting Graphs ϵ-far from $(k, c_{d,k}\mu\phi^2)$-clusterable.

We will show that Algorithm 2.1 rejects G with probability greater than $\frac{2}{3}$ if G is ϵ- far from $(2, \mu\phi^2\epsilon^2, c_d\mu^2\phi^4\epsilon^4)$-clusterable. We will do this by showing that if (u, v) is a 'representative' pair of vertices of G, then we need to perturb $\mathbf{q}_u$ and $\mathbf{q}_v$ by vectors of sufficiently large norm for $\mathbf{q}_u$ and $\mathbf{q}_v$ to lie on a line through the origin. This will imply that both of the eigenvalues of $A_{u,v}$ are large. We will also show that there a positive fraction of all pairs of vertices are representative and hence Algorithm 2.1 will find a representative pair with high probability.

We need the following inverse theorem from [2] which states that if G is ϵ-far from $(k, \phi_{in}, \phi_{out})$-clusterable then we can partition G into subsets of vertices that have sparse cuts between them.

Lemma 7. *[2, Lemma 4.5] If G is ϵ-far from $(k, \phi_{in}, \phi_{out})$-clusterable with $\phi_{in} \leq \alpha_{4.5}\epsilon$, then there exists a partition of V into $k+1$ subsets $V_1, \cdots, V_{k+1}$ such that for each $i, 1 \leq i \leq k+1, |V_i| = \Omega(\epsilon^2|V|/k)$ and $\phi_G(V_i) \leq c_{4.5}\phi_{in}\epsilon^{-2}$.*

Given Lemma 7 as a starting point, we now show that there are sufficiently many pairs of vertices (u, v) such that that $\mathbf{q}_u$ and $\mathbf{q}_v$ are far from being *'onedimensional.'* More precisely, we show that there is a positive fraction of pairs (u, v) such that if the origin, $\mathbf{q}_u + \mathbf{e}_u$ and $\mathbf{q}_v + \mathbf{e}_v$ are collinear, then we must have $\max(\|\mathbf{e}_u\|, \|\mathbf{e}_v\|)$ be sufficiently large. In the first step of our analysis which is Lemma 8, we borrow some of the tools from [7].

Lemma 8. *Let S_1 and S_2 be two disjoint subsets of vertices such that the cut $(S_i, V \setminus S_i)$ has conductance less than δ. Suppose that $|S_1| + |S_2| \leq \frac{2}{3}n$ and let $\mathbf{p}_{S_1}^t = \sum_{v\in S_1}\mathbf{p}_v^t$, $\mathbf{p}_{S_2}^t = \sum_{v\in S_2}\mathbf{p}_v^t, \mathbf{q}_{S_1}^t = \mathbf{p}_{S_1}^t - \frac{1}{n}\mathbf{1}$, and $\mathbf{q}_{S_2}^t = \mathbf{p}_{S_2}^t - \frac{1}{n}\mathbf{1}$. Let Π denote the projection onto the eigenvectors of M with eigenvalues greater than $1-2\delta$. Then if $|\alpha|+|\beta| \geq 1, |\alpha|, |\beta| \leq 1$, and $|\alpha+\beta| \leq 1$,*

$$\left\|\alpha\Pi\mathbf{q}_{S_1}^0 + \beta\Pi\mathbf{q}_{S_2}^0\right\|^2 \geq \frac{1}{12(|S_1|+|S_2|)}.$$

Proof. Let $s_1 = |S_1|, s_2 = |S_2|$. Let $\alpha \in \mathbb{R}$ and define the vector $\mathbf{f}$ as

$$f(v) = \begin{cases} \frac{\alpha}{s_1} & \text{if } v \in S_1 \\ \frac{\beta}{s_2} & \text{if } v \in S_2 \\ 0 & \text{otherwise} \end{cases}$$

and let $\mathbf{u} = \mathbf{f} - \frac{\alpha+\beta}{n}\mathbf{1} = \alpha\mathbf{q}^0_{S_1} + \beta\mathbf{q}^0_{S_2}$. Write $\mathbf{u}$ in the eigenbasis of $\mathbf{M}$ as $\mathbf{u} = \sum_i c_i\mathbf{v}_i$. We have

$$\begin{aligned} \|\mathbf{u}\|^2 = \sum_i \beta_i^2 &= \frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2} - \frac{(\alpha+\beta)^2}{n} \\ &\geq \frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2} - \frac{1}{n}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{u}^T\mathbf{L}\mathbf{u} &= \|\mathbf{u}\|^2 - \sum_i c_i^2\lambda_i \\ &= \sum_{i<j} M_{ij}(u_i - u_j)^2 \\ &\leq \frac{1}{2d}\frac{2\alpha^2}{s_1^2}\,\delta d s_1 + \frac{1}{2d}\frac{2\beta^2}{s_2^2}\,\delta d s_2 \\ &= \delta\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right). \end{aligned}$$

From above, it follows that

$$\sum_i c_i^2\lambda_i > \frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2} - \frac{1}{n} - \frac{\delta}{2}\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right).$$

Call $\lambda_i > 1 - 4\delta$ 'heavy.' Let H be the set of indices of the heavy eigenvalues. Letting $x = \sum_{i\in H}\beta_i^2$, we have

$$x + \left(\sum_i c_i^2 - x\right)(1 - 4\delta) > \frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2} - \frac{1}{n} - \frac{\delta}{2}\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right)$$

which implies that

$$x > \frac{3}{4}\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right) - \frac{1}{n}.$$

By Cauchy-Schwartz,

$$\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_1} \geq \frac{|\alpha| + |\beta|}{s_1 + s_2} \geq \frac{1}{s_1 + s_2}.$$

Now

$$\frac{2}{3(s_1+s_2)} \geq \frac{1}{n} \implies x \geq \frac{1}{12(s_1+s_2)}.$$

Hence,

$$\left\|\alpha\Pi\mathbf{q}^0_{S_1} + \beta\Pi\mathbf{q}^0_{S_2}\right\|^2 \geq \frac{1}{12(|S_1|+|S_2|)}.$$

□

We will now show that the conclusions of Lemma 8 also hold if we consider a large subset of S_1 and S_2.

Lemma 9. *Let S_1 and S_2 be two disjoint subsets of vertices that satisfy the conditions of Lemma 8. Let $T_i \subseteq S_i$ and $|T_i| = (1-\theta)|S_i|$ where θ is a sufficiently small constant for each i. Let $\mathbf{q}^t_{T_1} = \sum_{v\in T_1}\mathbf{q}^t_v$ and $\mathbf{q}^t_{T_2} = \sum_{v\in T_2}\mathbf{q}^t_v$. Furthermore, define α, β, and Π as in Lemma 8. Then*

$$\left\|\alpha\Pi\mathbf{q}^0_{T_1} + \beta\Pi\mathbf{q}^0_{T_2}\right\|^2 \geq \left(\frac{1}{\sqrt{12}} - 2\sqrt{\theta}\right)^2 \frac{1}{|S_1|+|S_2|}.$$

Proof. Let $\mathbf{q}^t_{S_1} = \sum_{v\in S_1}\mathbf{q}^t_v$ and $\mathbf{q}^t_{S_2} = \sum_{v\in S_2}\mathbf{q}^t_v$. Using the fact that $|T_i| = (1-\theta)|S_i|$ for each i, we can compute that

$$\begin{aligned}\left\|\alpha\mathbf{q}^0_{S_1} + \beta\mathbf{q}^0_{S_2} - \left(\alpha\mathbf{q}^0_{T_1} + \beta\mathbf{q}^0_{T_2}\right)\right\|^2 &= \frac{\theta}{1-\theta}\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right)\\ &\leq 2\theta\left(\frac{\alpha^2}{s_1} + \frac{\beta^2}{s_2}\right).\end{aligned}$$

Write $\alpha\mathbf{q}^t_{S_1} + \beta\mathbf{q}^t_{S_2} = \sum_i c_i\mathbf{v}_i$ and $\alpha\mathbf{q}^t_{T_1} + \beta\mathbf{q}^t_{T_2} = \sum_i w_i\mathbf{v}_i$ and let H denote the set of eigenvalues larger than $1-2\delta$ just like in Lemma 8. We have

$$\left\|\alpha\mathbf{q}^0_{S_1} + \beta\mathbf{q}^0_{S_2} - \alpha\mathbf{q}^0_{T_1} - \beta\mathbf{q}^0_{T_2}\right\|^2 \geq \sum_{i\in H}(\beta_i - w_i)^2.$$

Let $S = \frac{1}{s_1+s_2}$. From lemma 8 and the triangle inequality,

$$\begin{aligned}\sum_{i\in H} w_i^2 &> \left(\sqrt{\sum_{i\in H}\beta_i^2} - \sqrt{\sum_{i\in H}(\beta_i - w_i)^2}\right)^2\\ &> \left(\frac{\sqrt{S}}{\sqrt{12}} - 2\sqrt{\theta S}\right)^2\\ &= S\left(\frac{1}{\sqrt{12}} - 2\sqrt{\theta}\right)^2,\end{aligned}$$

as desired. □

Lemma 9 states that the line segment from $\mathbf{q}_{T_1}$ to $\mathbf{q}_{T_2}$ and the line segment from $\mathbf{q}_{T_1}$ to $-\mathbf{q}_{T_2}$ does not pass close to the origin. We now need a similar lemma that states that this happens for not only the averages $\mathbf{q}_{T_1}$ and $\mathbf{q}_{T_2}$, but also for a large number of *pairs of vertices*.

Lemma 10. *Let the sets S_1 and S_2 constant θ, and matrix Π satisfy the conditions of lemma 9. Then there are at least $\theta^2|S_1||S_2|$ pairs (u, v) where $u \in S_1, v \in S_2$, such that the following holds:*

$$\min_{0\le\alpha\le 1} \left\|\alpha\Pi\mathbf{q}_u^0 + (1-\alpha)\Pi\mathbf{q}_v^0\right\|_2^2 \ge \left(\frac{1}{24} - 2\theta - \frac{\sqrt{\theta}}{\sqrt{3}}\right)\frac{1}{|S_1|+|S_2|}. \tag{1}$$

Proof. Call a pair (u, v) *bad* if it does not satisfy (3) and *good* otherwise. Suppose for the sake of contradiction that there are more than $(1-\theta^2)|S_1||S_2|$ bad pairs. We will now show that there is a set $T \subseteq S_1$ where $|T| \ge (1-\theta)|S_1|$ such that for all $u \in T$, there are more than $(1-\theta)|S_2|$ vertices v in S_2 such that the pair (u, v) is bad. This must be true because otherwise, the number of bad pairs is at most

$$\theta|S_1||S_2| + (1-\theta)^2|S_1||S_2| < (1-\theta^2)|S_1||S_2|.$$

Hence, such a set T must exit. Now for every $u \in T$, let T_u denote the set of vertices in S_2 such that (u, v) is a bad for pair for all $v \in T_u$ and let $\mathbf{q}_{T_u}^t = \sum_{v\in T_u} \mathbf{q}_v^t$. We now claim that

$$\min_{0\le\alpha\le 1} \left\|\alpha\Pi\mathbf{q}_u^0 + (1-\alpha)\Pi\mathbf{q}_{T_u}^0\right\|_2^2 \le \left(\frac{1}{24} - 2\theta - \frac{\sqrt{\theta}}{\sqrt{3}}\right)\frac{1}{|S_1|+|S_2|}. \tag{2}$$

Suppose for the sake of contradiction that inequality (2) is not true. Consider the plane that passing through the origin $O, \Pi\mathbf{q}_u^t$ and $\Pi\mathbf{q}_{T_u}^t$ (see figure 1 for reference). If (2) does not hold, then the line segment L connecting Πq_u^t and $\Pi q_{T_u}^t$ lies entirely outside the circle C centered at the origin with radius equal to the right hand side of (2). By the hyperplane separation theorem, we know that there is a line l that separates L and C. Furthermore, we can guarantee that l never intersects C. Consider the two half spaces formed by l. From the definition of $\Pi\mathbf{q}_{T_u}$, we know that there is a point $v' \in T_u$ such that $\Pi\mathbf{q}_{v'}^t$ lies on the half space not containing C. Hence, the line segment from $\Pi\mathbf{q}_u^t$ and $\Pi\mathbf{q}_{v'}^t$ does not pass C. However, contradictions our assumption on the set T_u. Hence, the first inequality in (2) must hold. Now let $\mathbf{q}_{S_2}^t = \sum_{v\in S_2} q_v^t$. Letting $\beta = 1-\alpha$, we have

$$\begin{aligned}\left\|\alpha\Pi\mathbf{q}_u^0 + \beta\Pi\mathbf{q}_{S_2}^0\right\|_2 &= \left\|\alpha\Pi\mathbf{q}_u^0 + \beta\Pi\mathbf{q}_{T_u}^0 + \beta\Pi(\mathbf{q}_{S_2}^0 - q_{T_u}^0)\right\|_2 \\ &\le \left\|\alpha\Pi\mathbf{q}_u^0 + \beta\Pi\mathbf{q}_{T_u}^0\right\|_2 + |\beta|\left\|\mathbf{q}_{T_u}^0 - \mathbf{q}_{S_2}^0\right\|_2.\end{aligned}$$

We first bound the second term.

$$|\beta|\left\|\mathbf{p}_{S_2}^0 - \mathbf{p}_{T_u}^0\right\|_2^2 = \frac{\theta}{(1-\theta)|S_2|} \le \frac{2\theta}{|S_2|} \le \frac{4\theta}{|S_1|+|S_2|}$$

Then using (2), we have

$$\begin{aligned}\min_{0\le\alpha\le 1} \left\|\alpha\Pi\mathbf{q}_u^0 + \beta\Pi\mathbf{q}_{S_2}^0\right\|_2^2 &\le 2\left(\left(\frac{1}{\sqrt{24}} - \sqrt{2\theta}\right)^2\frac{1}{|S_1|+|S_2|} - \frac{4\theta}{|S_1|+|S_2|}\right) + \frac{8\theta}{|S_1|+|S_2|} \\ &= \left(\frac{1}{\sqrt{12}} - 2\sqrt{\theta}\right)^2\frac{1}{|S_1|+|S_2|}.\end{aligned}$$

Note that u was an arbitrary vertex in T. By letting $q_T^t = \sum_{u \in T} q_u^t$ and using the same geometric argument as above, we have that

$$\min_{0 \le \alpha \le 1} \left\| \alpha \Pi \mathbf{q}_T^0 + (1-\alpha) \Pi \mathbf{q}_{S_2}^0 \right\|_2^2 \le \left(\frac{1}{\sqrt{12}} - 2\sqrt{\theta} \right)^2 \frac{1}{|S_1| + |S_2|}.$$

However, this is a contradiction to lemma 9 so we are done. Hence, there must be at least $\theta^2 |S_1||S_2|$ pairs (u, v) where $u \in S_1, v \in S_2$, such that (3) holds. □

The conclusion of Lemma 10 also holds true if we let $\beta = \alpha - 1$. In this case, the argument is similar except we consider the line segment passing through $\Pi \mathbf{q}_u^t$ and $-\Pi \mathbf{q}_{T_u}^t$. Now we need to show that there are sufficiently many pairs (u, v) where $u \in S_1$ and $v \in S_2$ such that **both** the line segment from $\mathbf{q}_u$ to $\mathbf{q}_v$ and the line segment from $\mathbf{q}_u$ to $-\mathbf{q}_v$ does not pass close to the origin.

Lemma 11. *Let the sets S_1 and S_2 constant θ, and matrix Π satisfy the conditions of Lemma 9. Then there are at least $\theta^2 |S_1||S_2|$ pairs (u, v) where $u \in S_1, v \in S_2$, such that both the following hold:*

$$\begin{aligned} \min_{0 \le \alpha \le 1} \left\| \alpha \Pi \mathbf{q}_u^0 + (1-\alpha) \Pi \mathbf{q}_v^0 \right\|^2 &\ge \left(\frac{1}{24} - 2\theta - \frac{\sqrt{\theta}}{\sqrt{3}} \right) \frac{1}{|S_1| + |S_2|}, \\ \min_{0 \le \alpha \le 1} \left\| \alpha \Pi \mathbf{q}_u^0 + (\alpha - 1) \Pi \mathbf{q}_v^0 \right\|^2 &\ge \left(\frac{1}{24} - 2\theta - \frac{\sqrt{\theta}}{\sqrt{3}} \right) \frac{1}{|S_1| + |S_2|}. \end{aligned} \tag{3}$$

Proof. The proof of this lemma will be given in a future edition of the paper. □

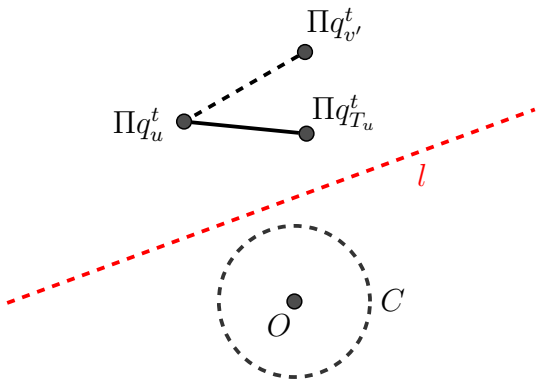

Figure 1: If the line segment connecting $\Pi \mathbf{q}_u^t$ and $\Pi \mathbf{q}_{T_u}^t$ does not intersect circle C then there must be a separating line H. Furthermore, there is a point $\Pi \mathbf{q}_{v'}^t$, where $v' \in T_u$, that lies on the opposite side of C with respect to H. Hence the segment connecting $\Pi \mathbf{q}_u^t$ and $\Pi \mathbf{q}_{v'}^t$ does not intersect C.

We now combine Lemmas 7 and 9. Furthermore, from our choice of constants in Section 3, we arrive at the following lemma.

Lemma 12. *Let $c = 512c_{4.5}c_{5.1}^2$. Call a pair (u, v) representative if both*

$$\min_{0\le\alpha\le 1} \left\|\alpha\mathbf{q}_u^t + (1-\alpha)\mathbf{q}_v^t\right\| = \Omega\left(\frac{1}{n^{1/2+c\mu}}\right)$$

and

$$\min_{0\le\alpha\le 1} \left\|\alpha\mathbf{q}_u^t + (1-\alpha)(-\mathbf{q}_v^t)\right\| = \Omega\left(\frac{1}{n^{1/2+c\mu}}\right).$$

A randomly chosen pair is representative with probability at least θ^2.

Lastly, we will show that if a pair (u, v) is representative, then we need to preturb $\mathbf{q}_u$ and $\mathbf{q}_v$ by vectors with large norm for the origin, $\mathbf{q}_1$, and $\mathbf{q}_2$ to be colliniear.

Lemma 13. *Suppose*

$$\min_{0\le\alpha\le 1} \|\alpha\mathbf{q}_1 + (1-\alpha)\mathbf{q}_2\| \ge R$$

and

$$\min_{0\le\alpha\le 1} \|\alpha\mathbf{q}_1 + (1-\alpha)(-\mathbf{q}_2)\| \ge R.$$

Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be such that the origin, $\mathbf{q}_1+\mathbf{e}_1$, and $\mathbf{q}_2+\mathbf{e}_2$ are collinear. Then, $\max(\|\mathbf{e}_u\|, \|\mathbf{e}_v\|) \ge R$.

Proof. See figure 2 for a reference. Consider the plane that passes through the origin $O, \mathbf{q}_1$, and $\mathbf{q}_2$. Let C denote the circle centered at the origin with radius R and let C' denote the circle centered at the origin that passes through $\mathbf{q}_1$. Let $\tilde{\mathbf{q}}_2$ denote the projection of q_2 onto C'. The first inequality in the lemma implies that the line segment connecting q_1 and q_2 does not intersect C. Then by considering the tangents from q_1 to C, we see that there is a sector S' on C', formed by the intersection of the tangents with C', such that $\tilde{\mathbf{q}}_2$ cannot lie on this sector. The angle of this sector is precisely $4\arcsin(R/\|\mathbf{q}_1\|) := 4\phi$. The second statement tells us that the line segment connecting $\mathbf{q}_1$ and $-\mathbf{q}_2$ does not intersect C. Hence, by reflecting $-\mathbf{q}_2$ across the origin, we see that there is a sector S'' identical to S' such that $\mathbf{q}_1$ is at the midpoint of this sector. Furthermore, $\tilde{\mathbf{q}}_2$ also cannot lie on this sector.

So far, we have shown that $\tilde{\mathbf{q}}_2$ cannot be close to diametrically opposite of q_1 and that $\mathbf{q}_1$ and $\tilde{q}_2$ cannot be close on C'. Now let L be an arbitrary line through the origin and let q' denote the intersection of L with C'. We claim that $\max(\angle\mathbf{q}_1O\mathbf{q}', \angle\tilde{\mathbf{q}}_2O\mathbf{q}') > \phi$. Indeed, without loss of generality, we can assume that $\tilde{\mathbf{q}}_2$ lies on the upper half circle with respect to $\mathbf{q}_1$, i.e., $0 \le \angle\mathbf{q}_1O\tilde{\mathbf{q}}_2 \le \pi$. We now consider two cases. If $\mathbf{q}'$ lies on the smaller arc connecting $\mathbf{q}_1$ and $\tilde{\mathbf{q}}_2$, then it is clear that the smallest $\max(\angle\mathbf{q}_1O\mathbf{q}', \angle\tilde{\mathbf{q}}_2O\mathbf{q}')$ can be is when $\tilde{\mathbf{q}}_2$ lies on the boundary of S'' which implies $\max(\angle\mathbf{q}_1O\mathbf{q}', \angle\tilde{\mathbf{q}}_2O\mathbf{q}') > \phi$.

In the case that $\mathbf{q}'$ lies on the larger arc connecting $\mathbf{q}_1$ and $\tilde{\mathbf{q}}_2$, it is also clear the smallest $\max(\angle\mathbf{q}_1O\mathbf{q}', \angle\tilde{\mathbf{q}}_2O\mathbf{q}')$ can be is when $\tilde{q}_2$ lies on the boundary of S'. In this case, it is also true that $\max(\angle\mathbf{q}_1O\mathbf{q}', \angle\tilde{\mathbf{q}}_2O\mathbf{q}') > \phi$. Thus, to project $\mathbf{q}_1$ and $\mathbf{q}_2$ to any line that passes through the origin, we must have that one of $\|\mathbf{e}_1\|, \|\mathbf{e}_2\| > \sin(\phi)\|\mathbf{q}_1\| = R$, as desired. □

Lemma 14. *If a pair (u, v) is representative then both of the eigenvalues of $\tilde{A}_{u,v}$ are larger than $\Omega\left(\frac{1}{n^{1/2+c\mu}}\right)$.*

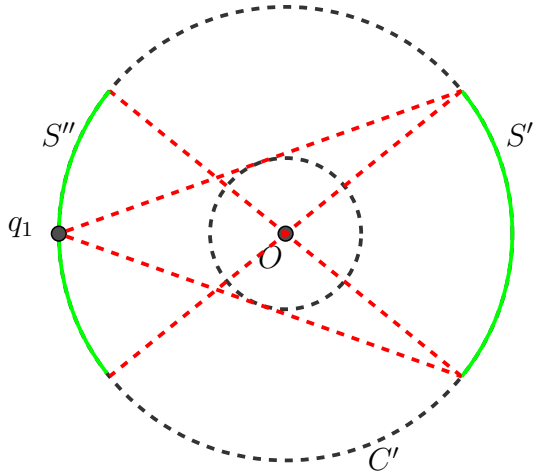

Figure 2: $\tilde{q}_2$ cannot like on the sectors S' and S''.

Proof. By Theorem 1.2, the eigenvalues of $\tilde{A}_{u,v}$ and $A_{u,v}$ differ by less than $\frac{2}{n^{1+\mu}}$. Hence it suffices to show that the eigenvalues of $A_{u,v}$ are larger than $\Omega\left(\frac{1}{n^{1/2+c\mu}}\right)$.

We know that $A_{u,v}$ is a gram matrix and hence positive semi-definite. Therefore, we can write $A_{u,v} = \kappa_1^2 w_1 w_1^T + \kappa_2^2 w_2 w_2^T$ where $\langle w_1, w_2 \rangle = 0$. Now let M_1 be a $n \times 2$ matrix with columns $\mathbf{q}_u$ and $\mathbf{q}_v$ and define $M_2 = \kappa_1 w_1 w_1^T + \kappa_2 w_2 w_2^T$. Denote the columns of M_2 as $\mathbf{q}'_u, \mathbf{q}'_v$. By the orthogonality of w_1 and w_2, we have that $M_1^T M_1 = M_2^T M_2$. Hence, $U = M_1 M_2^{-1}$ is an orthogonal matrix and the image of $\mathbf{q}'_u$ is $\mathbf{q}_u$ and similarly, the image of $\mathbf{q}'_v$ is $\mathbf{q}_v$. If $\kappa_2 < \Omega(1/n^{1/2+c\mu})$, then we can find $\mathbf{e}'_u, \mathbf{e}'_v$ such that $\|\mathbf{e}'_u\|, \|\mathbf{e}'_v\| < \kappa_2$ such that M_2 plus the matrix with $\mathbf{e}'_u, \mathbf{e}'_v$ in its columns is rank 1. Since U is an orthogonal matrix, we can thus find $\mathbf{e}_u, \mathbf{e}_v$ such that $\|\mathbf{e}_u\|, \|\mathbf{e}_v\| < \kappa_2$ and M_1 plus the matrix with columns $\mathbf{e}_u$ and $\mathbf{e}_v$ is rank 1. However, this contradicts Lemmas 12 and 13 so we are done. □

We finally show that Algorithm 2.1 passes the soundess case.

Lemma 15. *If G is ϵ-far from ϵ- far from $(2, \mu\phi^2\epsilon^2, c_d\mu^2\phi^4\epsilon^4)$-clusterable then Algorithm 2.1 rejects G with probability greater than $\frac{2}{3}$.*

Proof. A trial is rejected if we find a representative pair. Hence, the probability that a trial is rejected is at least $\frac{\theta^2|S_1||S_2|}{n^2}(1-\eta)^4 \geq \frac{\theta^2\epsilon^4}{64 \cdot 1152^2}$. Hence, the probability that all trials except is at most

$$\left(1 - \frac{\theta^2\epsilon^4}{64 \cdot 1152^2}\right)^S < \exp(-2) < \frac{1}{3}.$$

□

4 Future Work

The authors plan to release a future addition of this paper that gives a full proof of Lemma 9 and presents the analysis of Algorithm 2.1 for all values of k.

5 Acknowledgements

The authors would like to thank the MIT UROP+ program for the opportunity to work on this project. We will also like to thank the Raymond Stevens Fund for financially supporting this project.

References

[1] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics.

[2] Artur Czumaj, Pan Peng, and Christian Sohler. Testing cluster structure of graphs. *CoRR*, abs/1504.03294, 2015.

[3] Artur Czumaj and Christian Sohler. Testing expansion in bounded-degree graphs. *Combinatorics, Probability and Computing*, 19(5-6):693?709, 2010.

[4] Goldreich and Ron. Property testing in bounded degree graphs. *Algorithmica*, 32(2):302–343, Feb 2002.

[5] Oded Goldreich and Dana Ron. A sublinear bipartiteness tester for bounded degree graphs. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 289–298, New York, NY, USA, 1998. ACM.

[6] Oded Goldreich and Dana Ron. *On Testing Expansion in Bounded-Degree Graphs*, pages 68–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[7] Satyen Kale and C. Seshadhri. An expansion tester for bounded degree graphs. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming*, pages 527–538, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[8] Asaf Nachmias and Asaf Shapira. Testing the expansion of a graph. *Information and Computation*, 208(4):309 – 314, 2010.

[9] Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends in Theoretical Computer Science*, 5(2):73–205, 2010.

[10] Satu Elisa Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, August 2007.

[11] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, Dec 1912.