Testing Cluster Structure of Graphs

Artur Czumaj*
Dept. Computer Science
Centre for Discrete Mathematics
and its Applications (DIMAP)
University of Warwick
A.Czumaj@warwick.ac.uk

Pan Peng†
Dept. Computer Science
TU Dortmund &
State Key Lab of Computer
Science, Institute of Software,
Chinese Academy of Sciences
pan.peng@tu-dortmund.de

Christian Sohler‡
Dept. Computer Science
TU Dortmund
christian.sohler@tu-dortmund.de

ABSTRACT

We study the problem of recognizing the cluster structure of a graph in the framework of property testing in the bounded degree model. Given a parameter ε, a d-bounded degree graph is defined to be (k,ϕ)-*clusterable*, if it can be partitioned into no more than k parts, such that the (inner) conductance of the induced subgraph on each part is at least ϕ and the (outer) conductance of each part is at most $c_{d,k}\varepsilon^4\phi^2$, where $c_{d,k}$ depends only on d,k. Our main result is a sublinear algorithm with the running time $\widetilde{O}(\sqrt{n}\cdot\mathrm{poly}(\phi,k,1/\varepsilon))$ that takes as input a graph with maximum degree bounded by d, parameters k, ϕ, ε, and with probability at least $\frac{2}{3}$, accepts the graph if it is (k,ϕ)-clusterable and rejects the graph if it is ε-far from (k,ϕ^*)-clusterable for $\phi^* = c'_{d,k}\frac{\phi^2\varepsilon^4}{\log n}$, where $c'_{d,k}$ depends only on d,k. By the lower bound of $\Omega(\sqrt{n})$ on the number of queries needed for testing graph expansion, which corresponds to $k=1$ in our problem, our algorithm is asymptotically optimal up to polylogarithmic factors.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Theory

*Supported in part by Centre for Discrete Mathematics and its Applications (DIMAP) and by EPSRC grant EP/J021814/1.

†Supported by ERC grant No. 307696.

‡Supported by ERC grant No. 307696.

STOC'15, June 14–17, 2015, Portland, Oregon, USA.
ACM 978-1-4503-3536-2/15/06.
http://dx.doi.org/10.1145/2746539.2746618.

Keywords

graph property testing; graph clustering; random walks; spectral graph theory

1. INTRODUCTION

Cluster analysis is a fundamental task in data analysis that aims to partition a set of objects into maximal subsets (called *clusters*) of similar objects. In *graph clustering*, the objects to be clustered are the vertices of a graph and the edges of a graph describe relations between them. These relations may have interpretations for data analysis. For example, if the graph is the friendship graph of a social network, i.e., the vertices are the users of a social network and the edges correspond to friendship relations, edges may indicate that the users are socially related and/or have similar interests. In a co-author graph, where the vertices are authors and edges describe co-authorships, edges may be interpreted as a sign that the authors work in the same scientific community. A *cluster* is then a maximal subset of vertices that are *well-connected* to each other, where the precise meaning of being well-connected can be defined in various ways.

In many cases, once we know the interpretation of a single edge, there is a natural interpretation of clusters. For example, clusters in a friendship graph correspond to social groups or clusters in a co-author graph correspond to scientific communities. For similar reasons, a vast amount of graph clustering methods are applied to many different kinds of social/information/biological networks to reveal hidden cluster structure, etc. (see, e.g., surveys [11, 29, 33]).

Many efficient algorithms for finding clusters in a graph have been developed. However, with the increasing focus on the study of very large networks, we have to concentrate on new features of the clustering algorithms. For example, if one tries to find clusters in the World Wide Web or in a big social network, even linear time algorithms might be too slow. This is particularly important if one wants to study the temporal development of the clusters, which require to solve the problem on many instances (each for a different point of time). In such cases, we need *sublinear time algorithms*. We develop such an algorithm in this paper. Our algorithm can be used to test, if a given graph has a cluster structure, i.e., is composed of at most k clusters.

We will develop the algorithm in the framework of *Property Testing* for bounded degree graphs [15]. In this framework, an algorithm has oracle access to an undirected graph $G=(V,E)$ with a bound d on the maximum degree, with d

typically assumed to be constant. An algorithm is called a *property tester for a given property* Π (in our case, the property of all graphs that have a cluster structure with at most k clusters), if it accepts with probability at least $\frac{2}{3}$ every graph that has the property Π and rejects with probability at least $\frac{2}{3}$ every graph that is *ε-far from* Π. Here the notion of ε-far means that one has to change more than εdn edges to obtain a graph of maximum degree d that has property Π. If G is not ε-far from Π, then it is called *ε-close*. To give a property tester on a bounded degree graph G, we assume that G is given as an oracle, which allows us to perform *neighbor queries* to G such that for any input pair (v, i), the oracle returns the ith neighbor of vertex v if $i \leq d_G(v)$, and a special symbol if $i > d_G(v)$, where $d_G(v)$ is the degree of v. This framework of graph property testing was initiated by Goldreich and Ron [15]. In this model, it is known that several properties are testable in constant time, such as hyperfinite properties [24] (see also [7, 15] and the references therein). We now also know that properties such as bipartiteness [13] and expansion [9, 14, 17, 23] are testable in time $\widetilde{O}(\sqrt{n})$, with a nearly matching lower bound, and we need to perform at least $\Omega(n)$ queries to test 3-colorability [15]. For more results, see recent surveys [12, 31].

There are several ways to assess the cluster structure of a graph, such as k-means, cliques, modularity etc. One typically would want to argue that vertices in the same cluster should be well-connected and vertices from different clusters should be poorly-connected. In this paper, we use the concept of *conductance* to measure the quality of the cluster structure of a graph. Given a graph $G = (V, E)$ with maximum degree bounded by d, and a subset $S \subseteq V$, the *conductance of S* is defined as $\phi_G(S) := \frac{e(S, V \setminus S)}{d|S|}$, where $e(S, V \setminus S)$ denotes the number of edges coming out of S. Note that $\phi_G(V) = 0$. The *conductance of the graph G*, denoted as $\phi(G)$, is defined as the minimum conductance value over all possible subsets S of V with $|S| \leq |V|/2$. (For convenience, we define $\phi(G) = \frac{1}{d}$ if G is the singleton graph, that is, the graph consisting of a single isolated vertex with no edges.) For any $S \subseteq V$, let $G[S]$ be the induced subgraph of G on the vertex set S. Define the *inner conductance* of S to be the conductance of subgraph $G[S]$, namely, $\phi(G[S])$. To avoid confusion, we will also call the conductance $\phi_G(S)$ of S in G the *outer conductance* of S.

Kannan et al. [18] introduced conductance as a measure of the quality of a cluster and this notion has been later used in numerous more applied works (see, e.g., [33]). Further intuition has been employed to assert that a set S with small outer conductance has few connections to the outside of S, and a graph G with large conductance means that the vertices of G are well-connected with each other. Following this intuition, Oveis Gharan and Trevisan [27] and Zhu et al. [37] proposed to combine both outer conductance and inner conductance of a set S to measure whether S is a good cluster or not. That is, a set S is considered to be a good cluster if $\phi_G(S)$ is small and $\phi(G[S])$ is large. In [27], a graph G is defined to be clusterable if G can be partitioned into a number of disjoint parts so that each of them is such a good cluster. In this paper, we will use a related definition to characterize graphs with cluster structure.

1.1 Our results

We begin with the formal definition characterizing graphs with a cluster structure and state our main results. The following definition is inspired by the work of Oveis Gharan and Trevisan [27].

Definition 1. For a d-degree bounded undirected graph $G = (V, E)$ with n vertices and parameters k, ϕ, ε, we define G to be *(k, ϕ)-clusterable* if there exists a partition of V into h sets $C_1, \ldots, C_h$ such that $1 \leq h \leq k$, and for each i, $1 \leq i \leq h$, $\phi(G[C_i]) \geq \phi$ and $\phi_G(C_i) \leq c_{d,k}\varepsilon^4\phi^2$, where for fixed d, k, $c_{d,k}$ is a universal constant. We call each C_i a *ϕ-cluster* and the corresponding h-partition an *(h, ϕ)-clustering*.

The above definition formalizes the idea that the existence of an edge is an indicator that two vertices are similar, i.e., two persons are friends or two authors belong to the same scientific community, while the lack of an edge is a (weaker) sign of the opposite statement. Therefore, a cluster should be, intuitively, well-connected in the inside and poorly-connected to the outside. (We remark that the gap between the conductance of C_i and $G[C_i]$ in Definition 1 is a feature of our approach rather than an inherent property of the problem.)

In this paper, we develop an algorithm that with probability at least $\frac{2}{3}$, accepts every (k, ϕ)-clusterable graph and rejects every graph that is ε-far from every (k, ϕ^*)-clusterable graph, where $\phi^* = O_{d,k}(\frac{\phi^2\varepsilon^4}{\log n})$. (Throughout the paper we use the notation $O_{d,k}()$ to describe a function in the Big-Oh notation assuming that d and k are constant.) Our main result is that in sublinear time we can distinguish a clusterable graph from all graphs that are far from being clusterable.

THEOREM 1.1. *Let $c'_{d,k}$ be a suitable constant depending on d and k. There exists an algorithm that accepts every (k, ϕ)-clusterable graph of maximum degree at most d with probability at least $\frac{2}{3}$, 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 $\phi^* \leq c'_{d,k}\frac{\phi^2\varepsilon^4}{\log n}$. The running time of the algorithm is $\frac{\sqrt{n}}{\phi^2}(k \log n/\varepsilon)^{O(1)}$.*

One can question whether the gap between ϕ^* and ϕ in the form $\phi^* = O_{d,k}(\frac{\phi^2\varepsilon^4}{\log n})$ or similar is really required. We believe that for an algorithm with a somewhat similar time complexity, both the $\log n$ and the ε factors in the gap between ϕ and ϕ^* are necessary. For further discussion about this gap size we refer to Section 1.2.

Note also that in our results we allow for clusterings with at most k clusters (rather than with exactly k clusters). This can be justified by the fact that in the property testing framework, every (k, ϕ)-clusterable graph with exactly $h \leq k$ clusters is ε-close to some (k, ϕ^*)-clusterable graph with exactly k clusters, for any reasonable choice of parameters (one can simply remove all edges that are incident to $k - h$ vertices).

1.2 Comparison with testing expansion and discussion of the gap size

For $k = 1$, our problem is equivalent to that of *testing graph expansion*, the problem which has received significant attention in the past. Goldreich and Ron [14] were the first to study this problem in details and proved a lower bound $\Omega(\sqrt{n})$ on the number of queries for testing graph expansion in the bounded degree model. This result has been complemented by a proposed algorithm, which Goldreich and Ron

conjectured to be a property tester for the second largest eigenvalue (denoted by η_2) of the normalized adjacency matrix of a regularized version of the graph, in the sense that it accepts every graph with $\eta_2 \leq \eta$ and rejects every graph that is ε-far from having $\eta_2 \leq \eta^{\Theta(\mu)}$ for any $\mu > 0$. Note that by Cheeger's inequality, resolving of this conjecture would imply that the algorithm is also a property tester that accepts any graph with $\phi(G) \geq \phi$ and rejects every graph that is ε-far from being a ϕ^*-expander for $\phi^* = O(\mu\phi^2)$, where a graph G is called a ϕ-expander if $\phi(G) \geq \phi$. Czumaj and Sohler [9] proved a weaker version of this conjecture by showing that the algorithm from [14] can distinguish in time $\widetilde{O}(\sqrt{n})$ any ϕ-expander graph from graphs that are ε-far from being a ϕ^*-expander for $\phi^* = O(\frac{\phi^2}{\log n})$. Kale and Seshadhri [17] and Nachmias and Shapira [23] extended this result and proved that in $\widetilde{O}(n^{0.5+\mu})$ time the algorithm accepts graphs with expansion ϕ and reject graphs which are ε-far from having expansion $\phi^* = O(\mu\phi^2)$.

Since the best known methods require a gap between ϕ and ϕ^* already for the special case $k = 1$, it is clear that our work will also need a similar gap. It seems to be tempting to conjecture that — similarly to the case of testing expansion — it will suffice to reject (in the soundness) graphs that are ε-far from being $(k, \Theta(\mu\phi^2))$-clusterable for any $\mu > 0$, instead of having a $\log n$ factor dependency between ϕ and ϕ^*, as in our result. However, we do not think that this is possible and in the following we briefly sketch the differences from testing expansion and argue why the approach that led to a better gap for testing expansion is likely to fail (of course, this does not rule out other approaches, but this points to substantial obstacles to obtain an improved result).

Let u, v be any two vertices in the graph G, which for simplicity is now assumed to be d-regular and connected. Let λ_i be the i-th smallest eigenvalue and $\mathbf{v}_i$ be the corresponding eigenvector of the (normalized) Laplacian of G. It is known that the lazy random walk on G converges to the uniform distribution on its end-vertex. One can write (cf. the full version for details) the l_2^2-distance between the distribution $\mathbf{p}_v^\ell$ and $\mathbf{p}_u^\ell$ of the endpoints of the lazy random walks on G of length ℓ starting at v and u, respectively, as

$$\|\mathbf{p}_v^\ell - \mathbf{p}_u^\ell\|_2^2 = \sum_{i=1}^{n} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 (1 - \frac{\lambda_i}{2})^{2\ell} \ .$$

Since a lazy random walk on a regular graph converges to the uniform distribution, we have $\mathbf{v}_1(u) = \mathbf{v}_1(v) = 1/\sqrt{n}$. Therefore, in the case $k = 1$, by the fact that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 1$, we can upper bound $\|\mathbf{p}_v^\ell - \mathbf{p}_u^\ell\|_2^2$ by bounding the second smallest eigenvalue and by making a proper choice of the length of the walk ℓ.

If we want to extend this approach to $k > 1$, then our definition implies (cf. [19]) that in a (k, ϕ)-clusterable graph there is a significant gap between λ_h and λ_{h+1} for some h, $1 \leq h \leq k$, where h corresponds to the number of clusters in the instance. Now, assume for simplicity that $h = k$. Then we obtain that

$$\begin{aligned} \|\mathbf{p}_v^\ell - \mathbf{p}_u^\ell\|_2^2 &= \sum_{i=1}^{k} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 (1 - \frac{\lambda_i}{2})^{2\ell} \\ &\quad + \sum_{i=k+1}^{n} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 (1 - \frac{\lambda_i}{2})^{2\ell} \ . \end{aligned}$$

We can upper bound $\sum_{i=k+1}^{n} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 (1 - \frac{\lambda_i}{2})^{2\ell}$ by using the bound for λ_{h+1} in a similar way we can bound the entire term by bounding λ_2 in the case $k = 1$. However, the critical part is the first summand. It turns out that there are instances where the average l_2^2-distance between u, v from the same cluster is $\Omega(\frac{\phi^*}{d^3 n})$ for a certain reasonable choice of ℓ, such that the random walk mixes well in the cluster while does not escape from some non-expanding set containing the cluster too often (for more details, see discussions below and the full version of the paper). This seems to rule out an approach similar to [17, 23], as this approach requires a significantly smaller distance between $\mathbf{p}_v^\ell$ and $\mathbf{p}_u^\ell$.

1.3 Our techniques

We develop the first sublinear algorithm for testing if a graph is (k, ϕ)-clusterable, significantly extending earlier works on testing the expansion of a graph. Our algorithm draws a random sample set and tests for every pair of sample vertices if the distributions of the endpoints of a random walk starting at the two vertices are close in the l_2^2-distance. If this is the case, then it connects the two sample vertices by an edge in a *similarity graph*. At the end, the algorithm accepts the input graph if the similarity graph is a collection of at most k connected components.

Our main new contributions are as follows.

- Our algorithm is the first property tester that directly makes use of testing pairwise closeness of distributions induced by random walks. Previous related algorithms [9, 14, 16, 17, 23] tested if the distribution of the endpoints a random walk starting at a vertex v is close to the uniform distribution and then drew their conclusions about the structure of the graph. In our case, we do not know how the distribution looks like (it will be close to uniform *inside* every cluster, but this is not very helpful since the cluster is unknown to us and the support size of a distribution is hard to estimate [30]) and it may have significant distance from the uniform distribution.

- It is the first property tester that exploits (in the completeness case) a "somewhat stable" behaviour of the random walk distribution at a length where it is significantly different from the stationary distribution, i.e., we pick the length of the random walk in such a way that it is almost stable on its own cluster, and most of the probability mass will stay in some non-expanding set containing the cluster.

In order to test closeness of distribution, we use a recent tester for closeness of distributions in l_2-norm by Chan et al. [5], which gives slightly better bounds than the corresponding tester of Batu et al. [3]. A combination with a necessary condition on the l_2-norm of the distribution of the endpoints of the random walk from the sample vertices leads to improved bounds. It is tempting to think of this problem in the setting of l_1-norm since, for example, the distance between a random walk starting from different clusters is typically $\Omega(1)$ in l_1-norm. But this is misleading. It is known that no *stable* l_1-tester exists, i.e., l_1-testers cannot distinguish the case that distributions are close from the case that they are not [36] (l_1-testers can only distinguish between identical (or almost identical) distributions and distributions that are far away from each other). However, as already explained in

the previous section, we cannot hope for distributions to be arbitrarily close even if the random walks start in the same cluster. To address this difficulty, we will use the fact (noted earlier by Batu et al. [3]) that an l_1-tester can be reduced to an l_2-tester if the probability of every item is $O(n^{-1})$, which is likely to be the case if the graph is (k, ϕ)-clusterable.

We note that in the l_2^2-distance, a typical distance between the distribution of the endpoints of the random walks starting in two vertices from different clusters can be very small. For example, if we have two disconnected expanders (clusters) on $n/2$ vertices each, then for a sufficiently long random walk the distribution of the endpoints of the walk will be (almost) uniform on the cluster of the starting vertex. Therefore the distance between the distributions of the endpoints of random walks starting in different clusters will be $O(1/n)$. Furthermore, as we have argued above, the distance between the distributions of the endpoints of random walks will not be much smaller in the case that they come from the same cluster. Analyzing these two cases is one of the central technical challenges of our paper.

1.4 Other related work

In the context of property testing, Alon et al. [1] studied the problem of testing if a set of points in $\mathbb{R}^d$ is clusterable (see also [8]), but both their problem definition and techniques are quite different from ours. Kale et al. [16] gave a sublinear expansion reconstruction algorithm that outputs the neighborhood of any input vertex v in a $\Omega(\frac{\phi^2}{\log n})$-expander G' that is $\frac{\phi\varepsilon}{\log n}$-close to the input graph G, which is assumed to be ε-close to a ϕ-expander. In particular, they designed an algorithm that runs in $\widetilde{O}(\sqrt{n})$-time and distinguishes vertices from a large set that induces an expander from vertices that belong to a bad cut, by using uniform averaging random walks and testing if the distribution of endpoints of the walk is close to uniform distribution (in the l_1-norm distance) or not. This work does not (directly) compare distributions of the endpoints of the random walks starting from different vertices, as we do in our paper.

Our work is closely related to works on testing distributions. Batu et al. [2, 3] were the first to give sublinear time algorithms for testing the closeness of two discrete distributions and since then, a large body of work has been devoted to the problem of estimating the properties of distributions from a small number of samples (see the recent survey [32] and the reference therein). In particular, Levi et al. [20] gave an algorithm with complexity $\widetilde{O}(n^{2/3})$ to test whether a set of distributions over a domain of size n can be partitioned into k clusters. Very recently, Chan et al. [5] gave asymptotically optimal testers for the closeness of two distributions under both l_1 and l_2 settings.

Besides the related works in the literature of property testing, our work is also closely related to the area of graph partitioning and spectral clustering. Ng et al. [25] and Shi et al. [34] used the first few eigenvectors of some matrices to partition a graph (or a set of data) into sparsely connected clusters. Different ways of measuring clustering based on intra-cluster density vs. inter-cluster sparsity and some experimental results were given in [4]. Kannan et al. [18] proposed a bicriteria to measure the quality of a clustering, in which a good clustering is defined to be a partition of vertex sets such that each set in the partition has large inner conductance and few edges lying between different sets. They gave spectra based approximation algorithm for finding such a clustering. Lee et al. [19] and Louis et al. [21] recently gave theoretical analysis of some spectral algorithms that use the first k eigenvectors of the normalized Laplacian matrix for finding a k-partition of a graph such that each part is of small (outer) conductance, without any restriction on the inner conductance of the cluster. Zhu et al. [37, 26] gave personal PageRank based and flow based local algorithms for finding a set of large inner conductance and small outer conductance. Makarychev et al. [22] studied a semidefinite programming based algorithm in the semi-random model to find such a set. Tanaka [35] and Oveis Gharan and Trevisan [27] recently studied the existence and construction of a k-clustering such that each cluster is of large inner conductance and of small outer conductance, under the assumption that there is some gap between $\rho_G(k)$ and $\rho_G(k+1)$, where $\rho_G(k)$ is the minimum conductance of any k disjoint subsets of the graph (cf. Section 5). Dey et al. [10] considered the performance of a spectral clustering algorithm that applies a greedy algorithm for k-centers on some embedding induced by the first k eigenvectors of the graph Laplacian. Peng et al. [28] studied the eigenvector structures of the Laplacian of well-clustered graphs (which is very related to our definition of clusterable graphs) and the approximation ratio of k-means clustering algorithms on these graphs.

1.5 Organization of the paper

In Section 2, we give notations and definitions used throughout the paper. In Section 3, we give a formal description of our tester for clusterable graphs. We then present in Section 4 some central properties, which we use for proving our main result — Theorem 1.1. In Section 5, we show two spectral properties that will be used for proving our central properties. Section 6 has final conclusions. All missing proofs are deferred to the full version.

2. PRELIMINARIES

Let $G = (V, E)$ be an undirected and unweighted graph with maximum degree bounded by a constant d. Let $n := |V|$. For a vertex $v \in V$, let $d_G(v)$ be the degree of v. We assume that G is represented by its *adjacency list* and that we can access G through an oracle, which allows us to perform the *neighbor query* to G. That is, when the oracle is given as input a vertex v and an integer i, it outputs the i-th neighbor of v if $d_G(v) \geq i$, and a special symbol otherwise (in constant time).

As mentioned in the introduction, we will use Definition 1 of (k, ϕ)-clusterable graphs and ϕ-clusters inspired by [27] to characterize the cluster structure of graphs and the clusters therein. Note that a $(1, \phi)$-clusterable graph is an expander graph with conductance ϕ, which we abbreviate as ϕ-expander (this should not be confused with ϕ-cluster).

We are interested in testing if a given graph is (k, ϕ)-clusterable in sublinear time in the framework of property testing. Formally speaking, we will study the following problem: given parameters k, ϕ, ε, and a d-degree bounded graph G, we want to test if G is (k, ϕ)-clusterable or ε-far from being (k, ϕ^*)-clusterable with as few queries as possible, for ϕ^* being as close to ϕ as possible. We have the following definition of graphs that are ε-far from clusterable graphs.

Definition 2. A graph G (of maximum degree at most d) is *ε-far from (k, ϕ)-clusterable* if we have to add or delete

more than εdn edges to obtain a (k,ϕ)-clusterable graph of maximum degree at most d. If G is not ε-far from (k,ϕ)-clusterable then it is ε-*close* to (k,ϕ)-clusterable.

3. THE ALGORITHM

In this section, we describe our algorithm used in Theorem 1.1. We first introduce the following *random walk* on a d-bounded degree graph G that will be used in our algorithm. In this walk, if we are currently at vertex v, then in the next step, we choose randomly an incident edge (v,u) with probability $\frac{1}{2d}$ and move to u. With the remaining probability, which is at least $\frac{1}{2}$, we stay at v. Note that if we let G_{reg} denote the weighted d-regular graph that is obtained from G by adding an appropriate number of half-weighted self-loops, then this random walk is exactly a *lazy random walk* on G_{reg}. We will let $\mathbf{p}_v^\ell$ denote the distribution of endpoints of such a random walk of length ℓ starting at v. Our testing algorithm is given as follows.

k-Cluster-Test (G, s, ℓ, σ, k)

1. Sample a set S of s vertices independently and uniformly at random from V.
2. For any $v \in S$, let $\mathbf{p}_v^\ell$ be the distribution of endpoints of random walk of length ℓ starting at v.
3. For any $v \in S$, test if $||\mathbf{p}_v^\ell||_2^2 > \sigma$; if so, then abort and reject.
4. For each pair $u, v \in S$: if l_2 distribution tester accepts that $\|\mathbf{p}_u^\ell - \mathbf{p}_v^\ell\|_2^2 \leq \frac{1}{4n}$, then add an edge (u,v) in "similarity graph" H on vertex set S.
5. If H is the union of at most k connected components, then accept; otherwise, reject.

If the graph is (k,ϕ)-clusterable then we will show that (for the right choice of parameters) the distributions of the endpoints of random walks will be close if they come from the same cluster. Furthermore, Step 3 tests a necessary condition for the efficient l_2 distribution tester that will be used in Step 4, i.e., $||\mathbf{p}_v^\ell||_2^2$ is small, which is satisfied for almost all vertices in a (k,ϕ)-clusterable graph. The small l_2^2-norm property of distributions can then be exploited in the testing for closeness of distributions in Step 4 to obtain a better running time.

3.1 Implementation of distribution testing

Our algorithm relies on an efficient tester for the l_2-closeness of two distributions $\mathbf{p}$ and $\mathbf{q}$. The tester used in Step 4 of **k-Cluster-Test** was recently proposed by Chan et al. [5] and is similar to the l_2 distance tester in [3] that uses the statistics of collisions in the sample sets from both distributions $\mathbf{p}, \mathbf{q}$. The following is a direct corollary of Theorem 1.2 from [5].

THEOREM 3.1. *Let $c_{3.1}$ be some constant such that $c_{3.1} \geq 1$. Let $\delta, \xi > 0$ and let $\boldsymbol{p}, \boldsymbol{q}$ be two distributions over a set of size n with $b \geq \max\{\|\boldsymbol{p}\|_2^2, \|\boldsymbol{q}\|_2^2\}$. Let $r \geq c_{3.1} \cdot \frac{\sqrt{b}}{\xi} \ln \frac{1}{\delta}$. There exists an algorithm, denoted by l_2-**Distribution-Test**, that takes as input r samples from each distribution $\boldsymbol{p}, \boldsymbol{q}$, and accepts the distributions if $\|\boldsymbol{p} - \boldsymbol{q}\|_2^2 \leq \xi$, and rejects the distributions if $\|\boldsymbol{p} - \boldsymbol{q}\|_2^2 \geq 4\xi$, with probability at least $1 - \delta$. The running time of the tester is linear in its sample size.*

We also need an efficient algorithm to estimate the l_2^2-norm of the probability distribution of the endpoints of a random walk in a graph. In Step 3 of our algorithm **k-Cluster-Test** we will use **l_2^2-norm tester**, the performance of which is guaranteed in the following lemma (the proof follows almost directly from the proof of Lemma 4.2 in [9] that in turn is built on Lemma 1 in [14]).

LEMMA 3.2. *Let $G = (V, E)$ with $|V| = n$. Let $v \in V$, $\sigma > 0$ and $r \geq 16\sqrt{n}$. Let $t \geq 1$ and let $\boldsymbol{p}_v^t$ be the probability distribution of the endpoints of a random walk of length t from v. There exists an algorithm, denoted by l_2^2-**norm tester**, that takes as input r samples from $\boldsymbol{p}_v^t$ and accepts the distribution if $\|\boldsymbol{p}_v^t\|_2^2 \leq \sigma/4$ and rejects the distribution if $\|\boldsymbol{p}_v^t\|_2^2 > \sigma$, with probability at least $1 - \frac{16\sqrt{n}}{r}$. The running time of the tester is linear in its sample size.*

4. ANALYSIS OF k-CLUSTER-TEST

We outline the proof of our main theorem, Theorem 1.1. Our techniques are based on two intuitions. The first intuition is that if two "typical" vertices u, v are *from the same large cluster*, then the distributions of the endpoints of two *sufficiently long* random walks starting at u, v, respectively, are close; and if u, v are *separated by a non-expanding cut*, then the distributions of the endpoints of two *not so long* random walks from u, v, respectively, are far away from each other. If this intuition holds, then we can reduce our problem to the problem of testing the closeness of two distributions, and then use the returned results to decide whether the distributions induced by the random walks from different sampled vertices can be divided into k groups or not. In particular, if our input graph G is (k,ϕ)-clusterable, then we can get at most k connected components in our "similarity graph" H. (Actually, as will be seen from our proof, sampled vertices from the same cluster form a clique in H.) On the other hand, if G is far from being (k,ϕ^*)-clusterable, then we expect that we can get at least $k+1$ connected components in H . The latter is based on our second intuition that if G is far from being (k,ϕ^*)-clusterable, then there are at least $k+1$ (large) well separated sparse cuts. We present several lemmas that formalize these intuitions in Section 4.1 and then give the proof of Theorem 1.1 in Section 4.2.

4.1 Key properties

In this section, we state several lemmas describing the properties used in our analysis of **k-Cluster-Test**.

In the following we will formally state these key properties under the definition of a more general class of clusterable graphs, even though our main focus is on the study of properties of (k,ϕ)-clusterable graphs. To study detailed properties of (k,ϕ)-clusterable graphs and their dependencies on all parameters, we will use the following, more general definition of (k,ϕ_{in},ϕ_{out})-clusterable graphs, which follows the framework from [27].

Definition 3. For an undirected graph G, and parameters k, ϕ_{in}, ϕ_{out}, we define G to be (k,ϕ_{in},ϕ_{out})-*clusterable* if there exists a partition of V into h subsets $C_1, \ldots, C_h$ such

that $1 \leq h \leq k$ and for each i, $1 \leq i \leq h$, $\phi(G[C_i]) \geq \phi_{in}$, $\phi_G(C_i) \leq \phi_{out}$. We call each C_i a (ϕ_{in}, ϕ_{out})-*cluster* and the corresponding h-partition an $(h, \phi_{in}, \phi_{out})$-*clustering*.

We can define a graph G to be ε-far from $(k, \phi_{in}, \phi_{out})$-clusterable similarly to Definition 2. Note that a (k, ϕ)-clusterable graph from Definition 1 is exactly a $(k, \phi, c_{d,k}\varepsilon^4\phi^2)$-clusterable graph from Definition 3.

We first show that if the graph is $(k, \phi_{in}, \phi_{out})$-clusterable then for any large cluster C with $\phi(G[C]) \geq \phi_{in}$, there exists a large subgraph $\widetilde{C}$ such that the distributions of the endpoints of two random walks of length large enough starting from any two vertices $u, v \in \widetilde{C}$ are close in the l_2-norm (that is, the l_2 distance between $\mathbf{p}_u^\ell$ and $\mathbf{p}_v^\ell$ is small). The proof of this result relies on spectral properties of clusterable graphs given in Section 5.

LEMMA 4.1. *Let $0 < \alpha, \beta < \frac{1}{2}$. If graph $G = (V, E)$ is $(k, \phi_{in}, \phi_{out})$-clusterable, and $C \subseteq V$ is any subset such that $|C| \geq \beta n$ and $\phi(G[C]) \geq \phi_{in}$, then there exists $\alpha_{4.1} = \alpha_{4.1}(k, \alpha, \beta, d)$ and a universal constant $c_{4.1} > 0$ such that for any $t \geq \frac{c_{4.1}k^4 \log n}{\phi_{in}^2}$, $\phi_{out} \leq \alpha_{4.1}\phi_{in}^2$, there exists a subset $\widetilde{C} \subseteq C$ with $|\widetilde{C}| \geq (1-\alpha)|C|$ such that for any $u, v \in \widetilde{C}$, the following holds:*

$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \leq \frac{1}{4n} \ .$$

In order to use an efficient distribution tester (e.g., as the one given in Theorem 3.1), we need to guarantee that for a large fraction of vertices a sufficiently long random walk starting from a typical vertex will induce a distribution of its endpoints with small l_2-norms. We will prove the following lemma using spectral analysis of clusterable graphs.

LEMMA 4.2. *Let $0 < \alpha < 1$. If G is $(k, \phi_{in}, \phi_{out})$-clusterable, then there exists $V' \subseteq V$ with $|V'| \geq (1-\alpha)|V|$ such that for any $u \in V'$ and any $t \geq \frac{c_{4.2}k^4 \log n}{\phi_{in}^2}$, for some universal constant $c_{4.2} > 0$, the following holds:*

$$\|\boldsymbol{p}_u^t\|_2^2 \leq \frac{2k}{\alpha n} \ .$$

Note that the above lemma does not require any assumption about ϕ_{out}, and thus applies directly to any (k, ϕ)-clusterable graphs by substituting ϕ for ϕ_{in} in the lemma.

For the soundness of our algorithm, we need the following lemma that shows that given two well separated sets $A, B \subseteq V$, for any two "typical" vertices $u \in A$, $v \in B$, the l_2-norm of the difference between the corresponding distributions of endpoints of random walks of short length starting from u, v will be large. Our proof relies on the fact that any set A with small outer conductance has a large subset $\widehat{A}$ such that the random walk starting from any vertex in $\widehat{A}$ will stay inside A for a relatively long time.

LEMMA 4.3. *Let α and ψ be arbitrary with $0 < \alpha, \psi < 1$. Let $A \subseteq V$ be any subset of G such that $\phi_G(A) \leq \psi$. Then for any $t \geq 1$, there exists a subset $\widehat{A} \subseteq A$ with $|\widehat{A}| \geq (1-\alpha)|A|$ such that for any $v \in \widehat{A}$, the probability that the random walk of length t starting from vertex v never leaves A in all t steps is at least $1 - \frac{t\psi}{2\alpha}$.*

Furthermore, for any t, $1 \leq t \leq \frac{\alpha}{2\psi}$, any two disjoint subsets $A, B \subseteq V$ with $\phi_G(A), \phi_G(B) \leq \psi$, and any two vertices u, v such that $u \in \widehat{A}, v \in \widehat{B}$, the following holds:

$$\|\boldsymbol{p}_u^t - \boldsymbol{p}_v^t\|_2^2 \geq \frac{1}{n} \ .$$

Remark. We note that the above lower bound is almost tight up to constants. Consider the graph that is composed of two disconnected parts such that each of them is a ϕ_{in}-expanders of size $n/2$. Then for any two starting vertices u, v from two different parts, for $t = \Theta(\frac{\log n}{\phi_{in}^2})$, both $\mathbf{p}_u^t$ and $\mathbf{p}_v^t$ will be very close to the uniform distribution on each cluster, and therefore, the l_2^2 distance between these two distributions will be $O(1/n)$.

For the analysis showing that graphs far from clusterable will be rejected, we will use a property that if a graph $G = (V, E)$ is ε-far from any $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable graph, then its vertex set V can be partitioned into $k+1$ subsets $V_1, \ldots, V_{k+1}$, each of linear size and of small outer conductance.

LEMMA 4.4. *Let $\alpha_{4.4} = \alpha_{4.4}(d, k)$ be a certain constant that depends on d and k. If $G = (V, E)$ is ε-far from $(k, \phi_{in}^*, \phi_{out}^*)$-clusterable with $\phi_{in}^* \leq \alpha_{4.4} \cdot \varepsilon$, then there exist a partition of V into $k+1$ subsets $V_1, \ldots, V_{k+1}$ such that for each i, $1 \leq i \leq k+1$, $|V_i| \geq \frac{1}{1152k}\varepsilon^2|V|$ and $\phi_G(V_i) \leq c_{4.4}\phi_{in}^*\varepsilon^{-2}$, for some constant $c_{4.4} = c_{4.4}(d, k)$ and for any $0 \leq \phi_{out}^* \leq 1$.*

4.2 Proof of main result — Theorem 1.1

We will use Lemmas 4.1–4.4 to prove our main result — Theorem 1.1. In the rest of this section, we prove the completeness, soundness and analyze the running time of the tester k-**Cluster-Test**.

In the algorithm k-**Cluster-Test**, we set $s = \frac{1536k\ln(8(k+1))}{\varepsilon^2}$, $\ell = \frac{\max\{c_{4.1}, c_{4.2}\}\cdot k^4 \log n}{\phi^2}$, $\sigma = \frac{192sk}{n}$. In Theorem 3.1, we set $r = 192c_{3.1}\ s\sqrt{skn}\ln s = O(\frac{k^2(\ln k/\varepsilon)^{5/2}\sqrt{n}}{\varepsilon^3})$, $b = \frac{216sk}{n}$, $\xi = \frac{1}{4n}$, $\delta = \frac{1}{12s^2}$. In Lemma 3.2, we set $r = 192c_{3.1}s\sqrt{skn}\ln s$ and $\sigma = \frac{192sk}{n}$.

We specify now the constant $c_{d,k}$ that we used in the definition of a ϕ-cluster to be $c_{d,k} = \frac{\alpha_{4.1}(k, \frac{1}{24s}, \frac{1}{24ks}, d)}{\varepsilon^4} = \frac{c}{k^5 d^4 \ln^2(8(k+1))}$ for a universal constant c.

4.2.1 Completeness

We begin with showing that the algorithm k-**Cluster-Test** will accept k-clusterable graphs.

LEMMA 4.5. *If the input graph G is (k, ϕ)-clusterable, then with probability at least $\frac{2}{3}$, the algorithm k-**Cluster-Test** accepts G.*

PROOF. As indicated in the algorithm, we consider random walks of length ℓ. We apply Lemmas 4.1 and 4.2 to the (k, ϕ)-clusterable graph G, and we set $\phi_{in} = \phi$, $\phi_{out} = c_{d,k}\varepsilon^4\phi^2$, $t = \ell$, $\alpha = \frac{1}{24s}$, and $\beta = \frac{1}{24ks}$ in the lemmas. Note that by our definition of ϕ-cluster, the outer conductance of the cluster is at most $c_{d,k}\varepsilon^4\phi^2 \leq \alpha_{4.1}\phi^2$, since $c_{d,k}\varepsilon^4 = \alpha_{4.1}(k, \frac{1}{24s}, \frac{1}{24ks}, d)$, which implies that the conditions of Lemma 4.1 are satisfied for any ϕ-cluster of size at least βn in G. Since $\ell = \frac{\max\{c_{4.1}, c_{4.2}\}\cdot k^4 \log n}{\phi_{in}^2}$, we know that the chosen parameters meet all the preconditions in these lemmas.

Since G is (k,ϕ)-clusterable, there exists some h, $1 \le h \le k$, and a partition of the vertex set of G into h subsets $C_1,\ldots,C_h$, such that for every i, $1 \le i \le h$, we have $\phi(G[C_i]) \ge \phi$ and $\phi_G(C_i) \le c_{d,k}\varepsilon^4\phi^2$. For any vertex v, define $C(v)$ to be the unique cluster C_i to which v belongs.

We call a vertex v *good* if the following three conditions are satisfied:

1. $\|\mathbf{p}_v^\ell\|_2^2 \le \frac{48sk}{n}$.
2. $|C(v)| \ge \frac{1}{24ks}n$.
3. $v \in \widetilde{C(v)}$, where $\widetilde{C(v)} \subseteq C(v)$ is defined as in Lemma 4.1 by setting $C = C(v)$.

The success probability of the algorithm depends on the random coins of sampling and random walks. We show that with probability at least $\frac{7}{8}$ (over random coins of sampling), all vertices in the sample set S are good; and if all these vertices are good, then our tester will accept with probability at least $\frac{5}{6}$ (over random coins of random walks). Together, this means that with probability at least $\frac{7}{8}\cdot\frac{5}{6} = \frac{35}{48} \ge \frac{2}{3}$ the tester will accept. This will conclude the proof of the lemma.

CLAIM 4.6. *With probability at least $\frac{7}{8}$, all vertices in the sampled set S are good.*

PROOF. Let v be any vertex that is sampled uniformly at random from V. By Lemma 4.2, the probability that $\|\mathbf{p}_v^\ell\|_2^2 > \frac{48sk}{n}$ is at most $\alpha = \frac{1}{24s}$. Since there are at most k clusters, the probability that v belongs to a cluster of size at most $\frac{1}{24ks}n$ is at most the probability that v is one of at most $k\cdot\frac{n}{24ks}$ vertices in these small clusters, which is $\frac{1}{24s}$. In addition, since $|\widetilde{C(v)}| \ge (1-\alpha)|C(v)|$, the probability that $v \notin \widetilde{C(v)}$ is at most $\alpha = \frac{1}{24s}$. Overall, the probability that v is not good is at most $\frac{1}{24s}+\frac{1}{24s}+\frac{1}{24s} = \frac{1}{8s}$. By the above analysis and the union bound, with probability at least $1-\frac{1}{8s}\cdot s = \frac{7}{8}$, all sampled vertices in S are good. □

CLAIM 4.7. *Conditioned on the event that all the sampled vertices $v \in S$ are good, our tester will accept G with probability at least $\frac{5}{6}$.*

PROOF. Let $v \in S$. Since v is good, then $\|\mathbf{p}_v^\ell\|_2^2 \le \frac{48sk}{n} = \frac{\sigma}{4}$. Now by Lemma 3.2, **$l_2^2$-norm estimator** will reject v with probability at most $\frac{16\sqrt{n}}{r} \le \frac{1}{12s}$. By the union bound, the probability that we get rejected at step 3 of the algorithm is at most $\frac{1}{12}$.

For any two vertices u,v from S, if u,v belong to the same large cluster, then by Conditions 2–3 of good vertices and by Lemma 4.1, $\|\mathbf{p}_u^\ell - \mathbf{p}_v^\ell\|_2^2 \le \frac{1}{4n}$. Now recall that we have set $b = \frac{216sk}{n}, \xi = \frac{1}{4n}, \delta = \frac{1}{12s^2}$ and $r = 192c_{3.1}s\sqrt{skn}\ln s$ in Theorem 3.1. Then $b \ge \max\{\|\mathbf{p}_v^\ell\|_2^2, \|\mathbf{p}_u^\ell\|_2^2\}$, $r \ge c_{3.1}\cdot\frac{\sqrt{b}}{\xi}\ln\frac{1}{\delta}$, and we can ensure that with probability at least $1-\delta$, any call to l_2-**Distribution-Test** will accept the distributions $\mathbf{p}_u^t, \mathbf{p}_v^t$ if u,v belong to the same large cluster. By the union bound, the probability that there exist some call such that the distribution tester does not accept u,v if u,v are from the same cluster is at most $s^2\delta \le \frac{1}{12}$. Therefore, the probability that the algorithm does not reject at step 3 and all the calls to the l_2-**Distribution-Test** return the correct answer is at least $1-\frac{1}{12}-\frac{1}{12} = \frac{5}{6}$.

Now note that if for any $u,v \in S$ such that u,v belong to the same cluster, the distribution tester with input $\mathbf{p}_u^\ell, \mathbf{p}_v^\ell$ accepts, then there will an edge (u,v) in the "similarity graph" H. This further implies that all the vertices in S that are in the same cluster will form a clique. (But note that two sampled vertices from two different clusters might also be connected in H.) Since there are at most k clusters, we will get at most k connected components in H, and thus the tester will accepts G. □

4.2.2 Soundness

We present now a proof of the soundness of our tester.

LEMMA 4.8. *Let $\gamma = \gamma_{d,k} > 0$ be some constant depending on d,k. If the input graph $G=(V,E)$ is ε-far from (k,ϕ^*)-clusterable with $\phi^* \le \frac{\gamma\varepsilon^2}{s\ell}$, then the algorithm k-**Cluster-Test** rejects G with probability at least $\frac{2}{3}$.*

PROOF. We will use $\gamma = \min\{\frac{1}{48c_{4.4}}, \alpha_{4.4}\}$. Let us first observe that our choice of γ ensures that Lemma 4.4 implies the existence of a partition of V into $k+1$ disjoint sets $V_1,\ldots,V_{k+1}$ such that for each i, $1 \le i \le k+1$, $|V_i| \ge \kappa_1\varepsilon^2|V|$ and $\phi_G(V_i) \le \kappa_2\phi^*\varepsilon^{-2}$, for appropriate parameters $\kappa_1 = \frac{1}{1152k}$ and $\kappa_2 = c_{4.4}$.

Let $\alpha = \frac{1}{24s}$ (here α corresponds to the parameter α used in Lemma 4.3). For every set V_i, $1 \le i \le k+1$, let $\widehat{V}_i \subseteq V_i$ be the set of vertices $v \in V_i$ such that the probability that the random walk of length ℓ starting at v does not leave V_i is at least $1-\frac{\kappa_2\phi^*\ell}{2\alpha\varepsilon^2}$. We observe that since $\phi_G(V_i) \le \kappa_2\phi^*\varepsilon^{-2}$, we have $|\widehat{V}_i| \ge (1-\alpha)|V_i|$ by Lemma 4.3. Hence, our assumption that $|V_i| \ge \kappa_1\varepsilon^2|V|$ implies that $|\widehat{V}_i| \ge (1-\alpha)\kappa_1\varepsilon^2|V|$.

Let us call the sample set S chosen by the algorithm k-**Cluster-Test** to be *representative* if $\widehat{V}_i \cap S \ne \emptyset$ for every i, $1 \le i \le k+1$, and $S \subseteq \bigcup_{i=1}^{k+1}\widehat{V}_i$.

CLAIM 4.9. *The probability that the sample set S is representative is at least $\frac{5}{6}$.*

PROOF. For any set $X \subseteq V$, $\Pr[X \cap S = \emptyset] = (1-|X|/|V|)^s \le e^{-s|X|/|V|}$. Therefore, since $|\widehat{V}_i| \ge (1-\alpha)\kappa_1\varepsilon^2|V|$, the probability that S does not contain any element from $\widehat{V}_i$ is smaller than or equal to $e^{-s\widehat{V}_i/|V|} \le e^{-s(1-\alpha)\kappa_1\varepsilon^2}$. Hence, the union bound implies that the probability that there exists some $i \le k+1$ such that S does not contain any element from $\widehat{V}_i$ is at most $(k+1)\cdot e^{-s(1-\alpha)\kappa_1\varepsilon^2}$. In addition, the probability that there exists some vertex in S that belongs to $V \setminus (\bigcup_{i=1}^{k+1}\widehat{V}_i)$ is at most $s\cdot\alpha$. Therefore, the probability that S is representative is greater than or equal to $1-(k+1)\cdot e^{-s(1-\alpha)\kappa_1\varepsilon^2} - s\alpha$. Since $s = \frac{1536k\ln(8(k+1))}{\varepsilon^2}$ and $\alpha = \frac{1}{24s}$, we have $s(1-\alpha)\kappa_1\varepsilon^2 \ge \ln(8(k+1))$, and hence we can conclude that this probability is at least $\frac{5}{6}$. □

CLAIM 4.10. *If S is representative then the algorithm k-**Cluster-Test** rejects G with probability at least $\frac{5}{6}$.*

PROOF. Let $S_i := \widehat{V}_i \cap S$. Since S is representative, then $S = \bigcup_{i=1}^{k+1}S_i$. Recall that the algorithm k-**Cluster-Test** rejects G if one of the following two cases happen:

- there is a $v \in S$ such that **l_2^2-norm estimator** passes the testing of $\|\mathbf{p}_v^t\|_2^2 > \sigma$.

- for any $1 \le i < j \le k+1$, and any vertex pair u, v such that $u \in S_i$ and $v \in S_j$, (u, v) is not an edge in the "similarity graph" (because in that case the resulting graph H could not be a union of at most k connected components).

If there exists some $v \in S$ with $\|\mathbf{p}_v^t\|_2^2 > \sigma$, then by Lemma 3.2, **$l_2^2$-norm tester** with rejects v with probability at least $1 - \frac{16\sqrt{n}}{r} > \frac{2}{3}$ and we are done. Therefore, we assume in the following that for every $v \in S$, $\|\mathbf{p}_v^t\|_2^2 < \sigma$. Let us now observe that the probability that the algorithm **k-Cluster-Test** would reject G is lower bounded by the probability that for any $1 \le i < j \le k+1$, and any vertex pair u, v such that $u \in S_i$ and $v \in S_j$, **l_2-Distribution-Test** rejects the distributions $\mathbf{p}_u^\ell, \mathbf{p}_v^\ell$.

Our definition of sets $\widehat{V_1}, \widehat{V_2}, \ldots, \widehat{V_{k+1}}$ and the assumption on ϕ_{in}^* (which implies that $\ell \le \frac{\alpha}{2\kappa_2 \phi^* \varepsilon^{-2}} \le \frac{\alpha}{2 \max_i \{\phi_G(V_i)\}}$) ensure that for any $1 \le i < j \le k+1$, and any vertex pair u, v such that $u \in S_i$ and $v \in S_j$, we can apply Lemma 4.3 to obtain $\|\mathbf{p}_u^\ell - \mathbf{p}_v^\ell\|_2^2 \ge \frac{1}{n}$. We know, by Theorem 3.1 and our choice of b, ξ, δ in that theorem, that for every such pair v_i, v_j, **l_2-Distribution-Test** will accept the distributions $\mathbf{p}_{v_i}^\ell, \mathbf{p}_{v_j}^\ell$ with probability at most δ. Therefore, the probability that there exists some vertex pair u, v such that $u \in S_i$, $v \in S_j$, $1 \le i < j \le k+1$ and (u, v) is selected as an edge in the "similarity graph" (which would mean that **l_2-Distribution-Test** will accept $\mathbf{p}_u^\ell, \mathbf{p}_v^\ell$) is at most $s^2 \cdot \delta$. Therefore we can conclude that the algorithm **k-Cluster-Test** rejects G with probability at least $1 - s^2 \cdot \delta \ge \frac{5}{6}$. □

Now, the proof of Lemma 4.8 follows directly from Claims 4.9 and 4.10.

We set $c' \frac{\phi^2 \varepsilon^4}{\log n} \le \frac{\gamma \varepsilon^2}{s\ell}$ in Theorem 1.1. By our choice of s and ℓ, we can find a constant $c' = c'_{d,k}$ that depends on d and k satisfying this condition, and we then require that $\phi^* \le c' \frac{\phi^2 \varepsilon^4}{\log n}$.

4.2.3 Running time

Now we analyze the running time of the algorithm **k-Cluster-Test**. First note that to sample from distributions $\mathbf{p}_v^\ell$ for any $v \in V$, we need to perform r random walks of length ℓ from v and the corresponding time is $O(\ell r)$. Note that each invocation of either distribution tester runs in time linearly in the number of samples, that is r. Since we sampled s vertices, invoked **l_2^2-norm tester** for each vertex in the sample set S, and invoked **l_2-Distribution-Test** for each vertex pair in S, we know that the total running time of the algorithm is $O(\ell s r + r s + s^2 \frac{\sqrt{b}}{\xi} \ln \frac{1}{\delta}) = O(\frac{\sqrt{n} k^7 (\ln k)^{7/2} \ln \frac{1}{\varepsilon} \ln n}{\phi_{in}^2 \varepsilon^5})$.

This completes the proof of Theorem 1.1, which follows directly from Lemmas 4.5 and 4.8 and our analysis of the running time given above.

5. SPECTRAL PROPERTIES OF CLUSTERABLE GRAPHS

We present in this section two spectral properties of clusterable graphs, which will be used for proving the foregoing central properties, i.e., Lemma 4.1 – 4.4. We first observe that it will be sufficient for us to consider weighted d-regular clusterable graphs. This is true since our algorithm actually performs the lazy random walk on the (virtual) weighted d-regularized version G_{reg} of the input d-bounded degree graph G. In addition, under our definition, for any set $S \subseteq V$, the outer conductance $\phi_G(S)$ and inner conductance $\phi(G[S])$ of S in G are the same as outer conductance $\phi_{G_{\text{reg}}}(S)$ and inner conductance $\phi(G_{\text{reg}}[S])$ of S in G_{reg}, respectively. For this reason, in the rest of this section, we will assume that G is a weighted d-regular graph.

The proofs of spectral properties of clusterable graphs rely on a recent high-order Cheeger inequality by Lee et al. [19]. To state the inequality, we first introduce some notations.

Let $\mathbf{A}$ denote the adjacency matrix of G. Let $\mathcal{L} = \mathbf{I} - \frac{1}{d}\mathbf{A}$ be the Laplacian matrix of G, where $\mathbf{I}$ is the identity matrix. Let λ_i be the ith smallest eigenvalue of the Laplacian matrix $\mathcal{L}$ and let $\mathbf{v}_i$ denote the corresponding (unit) eigenvector. Note that the probability transition matrix of the lazy random walk on G is $\mathbf{W} := \frac{\mathbf{I} + \frac{1}{d}\mathbf{A}}{2}$, and it is straightforward to see that $\{1 - \frac{\lambda_i}{2}\}_{1 \le i \le n}$ is the set of eigenvalues of $\mathbf{W}$ with corresponding eigenvectors $\{\mathbf{v}_i\}_{1 \le i \le n}$.

For a d-regular graph G, let $\rho_G(k)$ denote the minimum value of the maximum conductance over any possible k disjoint nonempty subsets. That is,

$$\rho_G(k) := \min_{\text{disjoint } S_1, \ldots, S_k} \max_{1 \le i \le k} \phi_G(S_i) \ .$$

Lee et al. [19] proved the following higher-order Cheeger's inequality.

THEOREM 5.1 ([19]). *For any weighted d-regular graph G and any $k \ge 2$, it holds that*

$$\lambda_k/2 \le \rho_G(k) \le c_{5.1} k^2 \sqrt{\lambda_k} \ ,$$

where $c_{5.1}$ is some universal constant.

Remark. Lee et al. actually proved a stronger version of the above theorem that applies to any weighted graph, by using a *volume*-based definition of conductance. The weaker version given by Theorem 5.1 will be enough for our application.

Now we are ready to state the spectral properties of clusterable graphs, which are given in the following two lemmas. The first lemma says that in a k-clusterable graph there is a large gap between λ_h and λ_{h+1} for some $h \le k$.

LEMMA 5.2. *If G is weighted d-regular and $(k, \phi_{in}, \phi_{out})$-clusterable, then there exists h, $1 \le h \le k$, such that $\lambda_i \le 2\phi_{out}$ for any $i \le h$, and $\lambda_i \ge \frac{\phi_{in}^2}{c_{5.1}^2 h^4}$ for any $i \ge h+1$.*

PROOF. Since G is $(k, \phi_{in}, \phi_{out})$-clusterable, then for some h, $1 \le h \le k$, there exists a partition of V into h sets $C_1, \ldots, C_h$, such that $\phi(G[C_i]) \ge \phi_{in}$ and $\phi_G(C_i) \le \phi_{out}$ for any $i \le h$. From the latter, we obtain that $\rho_G(h) \le \max_i \phi_G(C_i) \le \phi_{out}$ and then by Theorem 5.1, $\lambda_h \le 2\phi_{out}$, and thus for any $i \le h$, $\lambda_i \le \lambda_h \le 2\phi_{out}$.

Next, let us consider an arbitrary $(h+1)$-partitioning $P_1, \cdots, P_{h+1}$ of V. We note that there must be at least one set in the partition, say P_{i_0}, such that $|P_{i_0} \cap C_j| \le \frac{1}{2}|C_j|$ for every $1 \le j \le h$. This is true since otherwise, for every i, $1 \le i \le h+1$, each P_i would contain more than half of the vertices of some cluster, say $C_{\pi(i)}$, that is, $|P_i \cap C_{\pi(i)}| > \frac{1}{2}|C_{\pi(i)}|$. Then, since there are h clusters $C_1, \ldots, C_h$, by the pigeonhole principle there would have to exist two indices i and j, $1 \le i < j \le h+1$, such that $\pi(i) = \pi(j)$. This would

mean that each of P_i and P_j contain more than half of the vertices from the same cluster $C_{\pi(i)}$, which is a contradiction since P_i and P_j are disjoint. This proves the existence of the set P_{i_0}.

Let $P := P_{i_0}$. For every $1 \leq i \leq h$, let $B_i := P \cap C_i$. Since each cluster C_i has large inner conductance, namely $\phi(G[C_i]) \geq \phi_{in}$, and since $|B_i| \leq \frac{1}{2}|C_i|$, we have $e(B_i, C_i \setminus B_i) \geq \phi_{in} d|B_i|$ for every $1 \leq i \leq h$. Hence, $\phi_G(P) = \frac{e(P,V\setminus P)}{d|P|} \geq \frac{\sum_{i=1}^{h} e(B_i,C_i\setminus B_i)}{d\sum_{i=1}^{h}|B_i|} \geq \phi_{in}$, and thus $\rho_G(h+1) \geq \phi_{in}$. Therefore by Theorem 5.1, we have $\phi_{in} \leq \rho_G(h+1) \leq c_{5.1} h^2 \sqrt{\lambda_{h+1}}$, which yields $\lambda_{h+1} \geq \frac{\phi_{in}^2}{c_{5.1}^2 h^4}$. □

The second lemma states that in a k-clusterable graph, for any large cluster C, the average value of $(\mathbf{v}_i(u) - \mathbf{v}_i(v))^2$ over all $|C|^2$ vertex pairs $u, v \in C$ is as small as $\Theta_d(\frac{\phi_{out}}{|C|\phi_{in}^2})$, for any $i \leq h \leq k$.

LEMMA 5.3. *Let $G = (V, E)$ be a weighted d-regular graph that is $(k, \phi_{in}, \phi_{out})$-clusterable and let $C \subseteq V$ be any subset with $\phi(G[C]) \geq \phi_{in}$. Then there is h, $1 \leq h \leq k$ such that for every i, $1 \leq i \leq h$, the following holds:*

$$\frac{1}{|C|} \sum_{u,v \in C} (\boldsymbol{v}_i(u) - \boldsymbol{v}_i(v))^2 \leq \frac{8d^4\phi_{out}}{\phi_{in}^2} \ .$$

PROOF. Since G is $(k, \phi_{in}, \phi_{out})$-clusterable, by Lemma 5.2, there exists h, $1 \leq h \leq k$, such that $\lambda_{h+1} \geq \frac{\phi_{in}^2}{c_{5.1}^2 h^4}$ and $\lambda_i \leq 2\phi_{out}$ for any $1 \leq i \leq h$. Hence, for any $i \leq h$, by the variational principle of eigenvalues, we have

$$\lambda_i = \frac{\sum_{(u,v)\in E}(\mathbf{v}_i(u) - \mathbf{v}_i(v))^2}{d} \leq 2\phi_{out} \ . \tag{1}$$

Let us recall a known result (see, e.g., [6, (1.5), p. 5]) that for any weighted graph $H = (V_H, E_H)$,[1]

$$\lambda_2(H) = \min_f \left\{ \frac{2 \cdot \mathrm{vol}_H(V_H) \cdot \sum_{(u,v)\in E_H}(f(u) - f(v))^2}{\sum_{u,v\in V_H}(f(u) - f(v))^2 d_H(u) d_H(v)} \right\} , \tag{2}$$

where $\lambda_2(H)$ denotes the second smallest eigenvalue of the normalized Laplacian of H, the volume $\mathrm{vol}_H(S)$ of a set $S \subseteq V_H$ is the sum of degrees of vertices in S, that is, $\mathrm{vol}_H(S) := \sum_{v\in S} d_H(v)$.

Let us consider the induced subgraph $H := G[C]$ on C. Let $\phi_H^{\mathrm{vol}}(S) := \frac{e(S,H\setminus S)}{\mathrm{vol}_H(S)}$ and

$$\phi^{\mathrm{vol}}(H) := \min_{S:\mathrm{vol}_H(S)\leq \mathrm{vol}_H(V_H)/2} \frac{e(S, H \setminus S)}{\mathrm{vol}_H(S)}.$$

Since $\phi(H) \geq \phi_{in}$, then it is straightforward to see that $\phi^{\mathrm{vol}}(H) \geq \frac{\phi_{in}}{d}$. Cheeger's inequality yields $\lambda_2(H) \geq \frac{\phi_{in}^2}{2d^2}$. Therefore, if we apply this bound to inequality (2), then,

$$\frac{2 \cdot \mathrm{vol}_H(V_H) \cdot \sum_{(u,v)\in E_H}(\mathbf{v}_i(u) - \mathbf{v}_i(v))^2}{\sum_{u,v\in V_H}(\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 d_H(u) d_H(v)} \geq \lambda_2(H) \geq \frac{\phi_{in}^2}{2d^2} \ .$$

Combining this with the fact that

$$\sum_{(u,v)\in E_H} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 \leq \sum_{(u,v)\in E_G} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 \leq 2d\phi_{out},$$

where the last inequality follows from inequality (1), we have that

$$\sum_{u,v\in V_H} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 d_H(u) d_H(v) \leq \frac{8d^3 \mathrm{vol}_H(V_H)\phi_{out}}{\phi_{in}^2} \ .$$

Next, since $\phi(H) \geq \phi_{in} > 0$ implies that $d_H(u) \geq 1$ for any $u \in V_H$, and since the fact that for any $u \in V_H$, $d_H(u) \leq d$ yields $\mathrm{vol}_H(V_H) \leq d|V_H| = d|C|$, using the bound above we obtain:

$$\begin{aligned} \sum_{u,v\in V_H} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 &\leq \sum_{u,v\in V_H} (\mathbf{v}_i(u) - \mathbf{v}_i(v))^2 d_H(u) d_H(v) \\ &\leq \frac{8d^3 \mathrm{vol}_H(V_H)\phi_{out}}{\phi_{in}^2} \\ &\leq \frac{8d^4|C|\phi_{out}}{\phi_{in}^2} \ . \end{aligned}$$

The completes the proof of Lemma 5.3. □

Remark. In the full version we show that Lemma 5.3 is essentially tight for $k = 2$ and constant ϕ_{in}. We prove that there is a $(2, \phi_{in}, \phi_{out})$-clusterable graph G with clusters C_1, C_2 such that for at least one cluster, say C_1, the average value of $(\mathbf{v}_2(u) - \mathbf{v}_2(u))^2$ between vertices u, v from C_1 is $\Omega(\frac{\phi_{out}}{d^3|C_1|})$.

6. CONCLUSIONS

We presented the first study of testing the clusterability of a graph in the bounded degree model, where we used both the inner conductance and outer conductance of a set to measure the quality of a cluster [27]. Our main result is an asymptotically optimal (up to polylogarithmic factors) algorithm with running time $\widetilde{O}(\sqrt{n} \cdot \mathrm{poly}(d, k, \varepsilon))$ to test if a graph is (k, ϕ)-clusterable or is ε-far from (k, ϕ^*)-clusterable for $\phi^* = O_{d,k}(\frac{\phi^2\varepsilon^4}{\log n})$. Our tester uses new ideas of testing pairwise closeness of distributions of random walks starting from a pair of sample vertices and draws from that conclusions on the graph structure. One of the key techniques underlying our analysis is a new application of the recent results on higher order Cheeger inequalities [19].

For further research, one of the major open problem is to narrow the gap between ϕ and ϕ^*, or to prove that the current gap is almost optimal for any tester with similar running time. As we discussed in Section 1.2, fundamentally new ideas are needed here.

It would also be very interesting to gain deeper insights of the structure of graphs that are ε-far from (k, ϕ^*)-clusterable, that is, to improve Lemma 4.4. More specifically, is it possible to get rid of the dependency of ε of the upper bounds for inner and/or outer conductance in Lemma 4.4?

7. REFERENCES

[1] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. *SIAM Journal on Discrete Mathematics*, 16(3):393–417, 2003.

[2] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000.

[1] We remark that in [6], the summation in the denominator is over all unordered pairs of vertices, while in our context, the summation is over all possible $|V_H|^2$ vertex pairs. Therefore, a multiplicative factor 2 appears in the numerator in equation (2), compared with the form in [6, (1.5), p. 5].

[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4, 2013.

[4] U. Brandes, M. Gaertler, and D. Wagner. Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithmics*, 12, 2007.

[5] S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014.

[6] F. R. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[7] A. Czumaj, O. Goldreich, D. Ron, C. Seshadhri, A. Shapira, and C. Sohler. Finding cycles and trees in sublinear time. *Random Structures and Algorithms*, 45(2):139–184, Sept. 2014.

[8] A. Czumaj and C. Sohler. Abstract combinatorial programs and efficient property testers. *SIAM Journal on Computing*, 34(3):813–842, 2005.

[9] A. Czumaj and C. Sohler. Testing expansion in bounded-degree graphs. *Combinatorics, Probability and Computing*, 19(5-6):693–709, 2010.

[10] T. K. Dey, A. Rossi, and A. Sidiropoulos. Spectral concentration, robust k-center, and simple clustering. *arXiv preprint arXiv:1404.1008*, 2014.

[11] S. Fortunato. Community detection in graphs. *Physics Reports*, 486, 2010.

[12] O. Goldreich. Introduction to testing graph properties. In O. Goldreich, editor, *Property Testing — Current Research and Surveys*, pages 105–141. Springer Verlag, 2011.

[13] O. Goldreich and D. Ron. A sublinear bipartiteness tester for bounded degree graphs. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 289–298, 1998.

[14] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[15] O. Goldreich and D. Ron. Property testing in bounded degree graphs. *Algorithmica*, 32:302–343, 2002.

[16] S. Kale, Y. Peres, and C. Seshadhri. Noise tolerance of expanders and sublinear expansion reconstruction. *SIAM Journal on Computing*, 42(1):305–323, 1013.

[17] S. Kale and C. Seshadhri. An expansion tester for bounded degree graphs. *SIAM Journal on Computing*, 40(3):709–720, 2011.

[18] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

[19] J. R. Lee, S. Oveis Gharan, and L. Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1117–1130, 2012.

[20] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

[21] A. Louis, P. Raghavendra, P. Tetali, and S. Vempala. Many sparse cuts via higher eigenvalues. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1131–1140, 2012.

[22] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Proceedings of the 44th ACM Symposium on Theory of Computing (STOC)*, pages 367–384, 2012.

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

[24] I. Newman and C. Sohler. Every property of hyperfinite graphs is testable. *SIAM Journal on Computing*, 42(3):1095–1112, 2013.

[25] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2001.

[26] L. Orecchia and Z. A. Zhu. Flow-based algorithms for local graph clustering. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1267–1286, 2014.

[27] S. Oveis Gharan and L. Trevisan. Partitioning into expanders. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1256–1266, 2014.

[28] R. Peng, H. Sun, and L. Zanetti. Partitioning well-clustered graphs with k-means and heat kernel. *arXiv preprint arXiv:1411.2021*, 2014.

[29] M. A. Porter, J.-P. Onnela, and P. J. Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[30] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

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

[32] R. Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012.

[33] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[34] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[35] M. Tanaka. Multi-way expansion constants and partitions of a graph. *arXiv:1112.3434*, 2013.

[36] G. Valiant and P. Valiant. The power of linear estimators. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 403–412, 2011.

[37] Z. A. Zhu, S. Lattanzi, and V. Mirrokni. A local algorithm for finding well-connected clusters. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 396–404, 2013.