TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2025.9010180
Volume xx, Number x, xxxxxxx 20xx

Testing Clusterability on Labeled Graphs

Pengzhi Gao, Peng Zhang*.

Abstract: Graph clustering is a fundamental task in analyzing complex networks, with wide applications in social networks, bioinformatics, and recommendation systems. Traditional clustering methods often focus merely on structural connectivity, which neglect edge labels that provide critical semantic information. Many real-world networks are labeled with vertices and edges carrying additional information in the form of labels. These labels provide crucial insights into the underlying structure and function of the network. Despite the importance of labeled graphs, the problem of testing clusterability on such graphs has been paid limited attention from the viewpoint of property testing. This paper addresses the problem of testing clusterability on labeled graphs, extending the framework for bounded degree graphs. This paper introduces a novel framework for testing clusterability on labeled graphs, integrating both graph structures and label consistency constraints. We introduce a new definition of clusterability that incorporates label information and propose a sublinear-time algorithm to efficiently verify with high probability whether a labeled graph exhibits a high-quality clustering structure, where each cluster is of high internal conductance, low external conductance and contains as many labels as possible. The algorithm achieves currently best query complexity of $\tilde{O}(\sqrt{n})$ in expectation, ensuring scalability for large-scale graphs. Our work bridges the gap between traditional graph clustering and the emerging field of labeled graph analysis, providing a theoretical foundation for testing clusterability in labeled networks. Theoretical analysis shows that the tester reliably accepts clusterable graphs and rejects those ϵ-far from clusterability with high probability.

Key words: property testing; labeled graphs; graph clustering; sub-linear time algorithm.

1 Introduction

1.1 Motivation

Graph clustering is a fundamental task in analyzing complex networks, with broad applications in social networks, bioinformatics, and recommendation systems. Besides complex networks, labeled graph analysis is pivotal for artificial intelligence technologies like wireless electric vehicles (EVs). Ramesh et al.[1] demonstrated how edge-labeled graphs model communication constraints (e.g., 'charging station proximity' or 'traffic congestion') to optimize EV routing. Ignoring such labels will decline cluster-based resource allocation, so the analysis for label-aware clusterability framework is necessary. Traditional clustering methods often focus solely on structural connectivity, overlooking edge labels that convey critical semantic information. Many real-world networks are labeled, with vertices and edges carrying additional information. These labels offer crucial insights into the network's underlying structure and function. Despite the significance of labeled graphs, testing clusterability on such graphs has received limited attention in property testing. This paper introduces a novel framework for testing clusterability

• Pengzhi Gao and Peng Zhang are at School of Software, Shandong University, Jinan and 250101, China. E-mail: gaopengzhi@mail.sdu.edu.cn, algzhang@sdu.edu.cn
* Corresponding author.
Manuscript received: 2025-8-28; accepted: 2025-12-22

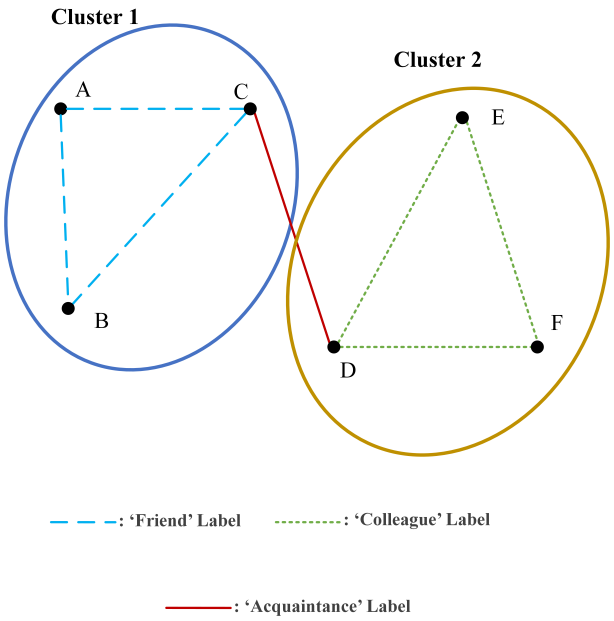

Fig. 1 A labeled graph that can be divided into two clusters.

on labeled graphs, integrating constraints for both structural topology and label consistency.

To the best of our knowledge, there is no enough research on semantic partitioning of edge labels on graphs. This paper addresses this gap by proposing a labeled graph clustering testing framework that combines the structural information of graphs and the labeling information of edges. For example, Fig. 1 consists of two main clusters and an outlier edge connecting them: Cluster 1 contains nodes A, B, and C, representing a tight social group where all connections are labeled as 'friend'. This cluster has high internal conductance (strong connections within the cluster) and label consistency (uniform edge labels). Cluster 2 includes nodes D, E, and F, representing a professional network with all edges labeled as 'colleague'. Similar to Cluster 1, it has good internal conductance and label consistency. The outlier edge connecting node C from Cluster 1 to node D from Cluster 2 has an 'acquaintance' label. This edge represents a weaker connection between the two clusters, contributing to the external conductance of each cluster.

1.2 Related Work

Research in graph clustering and property testing has yielded substantial advancements over recent decades. The foundational work by Goldreich and Ron[2,3] established key paradigms for property testing in bounded degree graphs, notably introducing a sublinear-time algorithm for testing bipartiteness, which remains highly influential in the field. Building upon this foundation, Czumaj et al.[4] made significant strides by developing algorithms specifically designed to test for cluster structure within graphs. Their contributions also included establishing important relationships between two distinct property testing models applicable to bounded degree directed graphs[5], further solidifying the theoretical framework. A pivotal theoretical breakthrough was achieved by Alon and Shapira[6], who demonstrated the testability of every monotone graph property, greatly expanding the scope of problems amenable to property testing techniques.

Recent advances integrate semantic features into clustering architectures. Chen et al.[7] leveraged edge labels as semantic constraints to refine node embeddings, achieving state-of-the-art accuracy in community detection. However, it lacks theoretical guarantees for clusterability verification. Similarly, constrained clustering methods proposed by Jia et al.[8] handled pairwise constraints but focused on vector data, not graph-structured labels. This theoretical progress naturally motivated more specific investigations into clusterability. In this vein, Kapralov et al.[9] established rigorous algorithms and provided fundamental lower bounds for testing graph clusterability, delineating the computational feasibility of this task. More recent developments continue to utilize these basic theories. Gluch et al.[10] for instance, addressed practical computational constraints by developing sublinear-time spectral clustering oracles, enabling efficient approximate analysis of large networks. Concurrently, Peng and Wang[11] refined the understanding of property testing models for directed graphs, investigating and establishing an optimal separation between two such models in the bounded-degree setting.

The study of graph clustering and property testing has progressively evolved to address increasingly complex scenarios, driving the development of more sophisticated algorithmic techniques. A key focus has been enhancing recognition efficiency for fundamental structural properties. Forster et al.[12] made significant strides in this direction by introducing algorithms capable of both computing and testing small vertex connectivity in near-linear time and query complexity, substantially improving practical feasibility for large-scale networks. Moving beyond specific connectivity

measures towards broader graph families, Kumar et al.[13] achieved a notable breakthrough for minor-closed properties within bounded degree graphs. They established the existence of a highly efficient tester requiring only poly$(d\varepsilon - 1)$ queries, demonstrating the testability of rich structural characterization. In the meanwhile, the exploration of clustering complexity itself has deepened. Li et al.[14] recently pioneered the examination of higher-order clusterability, venturing beyond traditional pairwise connections to analyze more intricate high-dimension structural patterns within graphs. Collectively, these advances improve efficiency for core problems, expand the scope of testable properties to complex graph classes, and probe richer notions of cluster structure. It highlights the sustained efforts to enhance graph clustering methodologies and extends the reach of property testing to increasingly complicated graph properties and structural paradigms.

1.3 Our Contributions

This paper addresses the problem of testing clusterability on labeled graphs, extending the framework proposed by Czumaj et al.[4] for bounded degree graphs. We introduce a new definition of clusterability that incorporates label information. We propose a sublinear-time algorithm to efficiently verify whether a labeled graph exhibits a high-quality clustering structure, defined by internal/external conductance and dominant label ratios within clusters. The algorithm achieves currently the best query complexity of $\tilde{O}(\sqrt{n})$ in expectation, ensuring scalability for large-scale graphs. Our work bridges the gap between traditional graph clustering and the emerging field of labeled graph analysis, providing a theoretical foundation for testing clusterability in labeled networks. Theoretical analysis demonstrates that the tester reliably accepts clusterable graphs and rejects those ε-far from clusterability with high probability.

The main contributions of this paper are as follows.

(i) A new problem framework is proposed for labeled graph clusterability testing, which integrates the structural information of the graph and the labeling information of the edges.

(ii) An efficient sublinear time algorithm (i.e., tester) is designed to quickly recognize high-quality clustering structures on large-scale labeled graph.

(iii) The time complexity of the tester is $\tilde{O}(\sqrt{n})$, which is currently the best time to test clusterability on labeled graphs.

2 Notations and Preliminaries

In this section, we detail the notation definitions and basic terminology to be used in this paper. Let the labeled graph be $G = (V, E, L)$, where V is the vertex set whose size is $|V| = n$; E is the edge set with maximum degree d; $L : E \to \Sigma$ is function for labeling, where Σ is set of all labels. Given a subset $C \subseteq V$, external conductance of C is defined as $\phi_G(C) = \frac{e(C,\, V\backslash C)}{d|C|}$, internal conductance of C is defined as $\phi(G[C]) = \min_{\substack{S\subset C \\ 0<|S|\leq|C|/2}} \frac{e(S,C\backslash S)}{d|S|}$, where $e(S,\, C\backslash S)$ denotes the set of edges connecting S and $C\backslash S$ in the induced subgraph $G[C]$ and $e(C,\, V\backslash C)$ denotes the set of edges and the two endpoints of each edge are in the set C and in the set $V\backslash C$.

Definition 2.1 (Label-aware clusterability) Given the parameters $k, \phi_{in}, \phi_{out}, \lambda_{in}$, graph G is (k, ϕ_{in}, ϕ_{out}, λ_{in})-clusterability if there exists a partition $C = \{C_1, \ldots, C_h\}$ $(h \leq k)$ that satisfies:

1 Internal conductance: $\forall i, \phi\left(G\left[C_i\right]\right) \geq \phi_{in}$;

2 External conductance: $\forall i, \phi_G\left(C_i\right) \leq \phi_{out}$;

3 (Label consistency conditions) $\forall i, \exists \sigma_i \in \Sigma, \lambda_i = \frac{|\{e\in C_i : L(e)=\sigma_i\}|}{d|C_i|} \geq \lambda_{in}$.

Definition 2.2 ($(k, \phi_{in}, \phi_{out}, \lambda_{in})$-clusterability tester) Given a neighbor query oracle to input graph $G = (V, E, L)$ with degree bound d and parameters k, ϕ_{in}, ϕ_{out}, λ_{in}, ε, we say that an algorithm is a tester for the $(k, \phi_{in}, \phi_{out}, \lambda_{in})$-clusterability if the algorithm satisfies the following acceptance property and rejection property.

(i) (*Acceptance*) If G has a clusterability structure that satisfies Definition 2.1, the algorithm returns acceptance with probability at least $\frac{2}{3}$.

(ii) (*Rejection*) if G is ε-far from being clusterability structure, the algorithm returns rejection with probability at least $\frac{2}{3}$. Here, by "ε-far" we mean that G must be modified (adding or deleting) at least εdn edges or labels so that it has the clusterability structure.

3 Algorithm of Testing Label-aware Clusterability

3.1 Design of k-clustering Tester

We propose a tester, i.e., the Label-aware Clusterability Tester, that aims to efficiently verify whether a labeled graph G satisfies the $(k, \phi_{in}, \phi_{out}, \lambda_{in})$-clusterability. Given threshold $\tau = \frac{\varepsilon^2}{16n}$ and $\widehat{C}_i$ denotes the estimated cluster corresponding to the i-th connectivity component in H. The tester is shown as Algorithm 1.

Algorithm 1: Label-aware Clusterability Tester

Input: Query oracle of labeled graph $G = (V, E, L)$, parameters $k, \phi_{in}, \phi_{out}, \lambda_{in}, \varepsilon$

Output: **Accept** or **reject**

1 Sample a set S of s vertices uniformly at random, where $s = \Theta\left(\frac{k \log n}{\varepsilon^2}\right)$;

2 For each vertex $v \in S$, perform $m = \Theta(\log n)$ lazy random walks of length $l = \Theta\left(\frac{\log n}{\phi_{in}^2}\right)$;

3 Construct a similarity graph H: For each $u, v \in S$, a vertex pair (u, v) is connected by an edge if and only if its random walk distribution distance is less than $\tau = \frac{\varepsilon^2}{16n}$;

4 For each connected subgraph $\widehat{C}_i$ (estimated cluster) of graph H, sample $s' = \Theta\left(\frac{\log(k/\delta)}{\lambda_{in}^2 \varepsilon^2}\right)$ edges uniformly at random;

5 Calculate the dominant label ratio $\widehat{\lambda}_i$ and verify if $\hat{\lambda}_i \leq \lambda_{in} + \frac{\varepsilon}{2}$;

6 If H contains at most k connectivity components, return **accept**. Otherwise, return **reject**.

The algorithm operates in two main phases, that is, connectivity estimation and label consistency verification, leveraging a construction of similarity graph and spectral graph analysis. To avoid exhaustive graph traversal, the algorithm uniformly samples $\widetilde{O}(\sqrt{n})$ vertices and edges, ensuring scalability for large graphs. Lazy random walk is a variant of the standard random walk. At each step, it has a fixed probability (typically 1/2) of staying at the current vertex rather than always moving to the neighbor vertex. Multi-step lazy random walks are employed to capture both local and global structural patterns. By analyzing the l_2^2-distance between endpoint distributions, the algorithm constructs a similarity graph H, where connectivity components reflect potential clusters. For each candidate cluster, Monte Carlo sampling estimates the dominant label ratio, ensuring label consistency within clusters. The total time complexity is $\tilde{O}(\sqrt{n} \cdot \mathrm{poly}(k, 1/\phi, 1/\epsilon))$.

3.2 Correctness of the k-clustering Tester

The correctness of Algorithm 1 hinges on rigorously bounding the probability of misclassification under both clusterability and ε-far from clusterability. To bridge the algorithmic steps with the formal proof of Theorem 3.6, we outline the proof strategy here, leveraging tools from spectral graph theory, concentration inequalities and property testing.

Central to the analysis is the interplay between clusters conductance and label consistence. For the acceptance case, we show that the similarity graph H preserves the ground-truth cluster partitioning with high probability, relying on the rapid mixing of lazy random walks within clusters (Lemma 3.1 and Lemma 3.2). The l_2^2-distance threshold τ is calibrated to ensure that vertices from the same cluster are connected in H, while inter-cluster pairs remain disconnected (Lemma 3.3). For the rejection case, we demonstrate that any ε-far graph induces $k+1$ disconnected components in H, derived via a combinatorial argument on minimally similar subgraphs (Lemma 3.4).

(Accept) To establish the correctness of Theorem 3.6, we leverage foundational results from spectral graph theory and random walk analysis. Specifically, the connection between conductance and mixing properties of lazy random walks plays a crucial role. Lemma 3.2 formalizes this relationship: for any cluster C_i with internal conductance ϕ_{in}, a sufficiently long lazy random walk (of length $l = \Theta\left(\frac{\log n}{\phi_{in}^2}\right)$) starting from a vertex within C_i will converge to a distribution close to the uniform distribution over C_i. This convergence guarantees that l_2^2-distance between the endpoint distributions of walks originating from the same cluster is bounded, thereby enabling the construction of a similarity graph H where intra-cluster vertices are densely connected. Crucially, the parameter choices for s (sample size) and l (walk length) in Lemma 3.1 are derived from the mixing time bounds in Lemma 3.2, ensuring that the structural properties of G are faithfully captured in H. This interplay between conductance-driven mixing and algorithmic sampling forms the backbone of our connectivity estimation framework.

The following Lemma 3.1 is about the similarity graph H. Lemma 3.1 comes from the work of Czumaj et al.[4]. Its proof is omitted here.

Lemma 3.1 Let $s = \Theta\left(\frac{k \log n}{\varepsilon^2}\right)$, $l = \Theta\left(\frac{\log n}{\phi_{in}^2}\right)$. The tester (Algorithm 1) satisfies the following two

properties.

(i) If G satisfies the clusterable condition, the similarity graph H can get at most k connected components with probability at least $\frac{5}{6}$;

(ii) If G is ε-far from being clusterable, the similarity graph H can get more than k connected components with probability at least $\frac{5}{6}$.

Lemma 3.2 Let G be a graph satisfying the following properties: there exists a k-partition $\{C_1, C_2, \ldots, C_k\}$ such that for each C_i, the internal conductance satisfies $\phi(G[C_i]) \geq \phi_{in}$, and the external conductance satisfies $\phi_G(C_i) \leq \phi_{out}$, where the parameters adhere to $\phi_{out} = O\left(\frac{\varepsilon^4 \phi_{in}^2}{\log n}\right)$. Suppose that we sample s vertices and perform l steps of lazy random walks from each sampled vertex. Then, the constructed similarity graph H satisfies the following properties with high probability.

(i) All vertices belonging to the same cluster C_i form a connected subgraph in H.

(ii) Vertices from distinct clusters remain disconnected in H.

Proof We use spectral graph theory and the properties of lazy random walks on graphs with high internal conductance. Let $G[C_i]$ be the induced subgraph on cluster C_i with internal conductance ϕ_{in}. Let π_{C_i} be the uniform distribution over C_i and $deg(v)$ be the degree of v. The mixing time of a lazy random walk is related to the spectral gap. Let λ_2 be the second largest eigenvalue of transition matrix of the lazy random walk on $G[C_i]$. By Cheeger's inequality[4], we have

$$1 - \lambda_2 \geq \frac{\phi_{in}^2}{2}.$$

The convergence of the lazy random walk is bounded by

$$\left\|p_v^l - \pi_{C_i}\right\|_2 \leq \sqrt{\frac{1}{\pi_{C_i}(v)}} \cdot (1-\lambda_2)^l.$$

Since $\pi_{C_i}(v) = \frac{1}{|C_i|}$ and $deg(v) \leq d$, we have

$$\frac{1}{\pi_{C_i}(v)} = |C_i|.$$

Thus,

$$\left\|p_v^l - \pi_{C_i}\right\|_2 \leq \sqrt{|C_i|} \cdot (1-\lambda_2)^l.$$

Using the bound $1 - \lambda_2 \leq e^{-\lambda_2} \leq e^{-(1-\lambda_2)}$, we get

$$(1-\lambda_2)^l \leq e^{-(1-\lambda_2)l} \leq e^{-\frac{\phi_{in}^2 l}{2}}.$$

Therefore,

$$\left\|p_v^l - \pi_{C_i}\right\|_2 \leq \sqrt{|C_i|} \cdot e^{-\frac{\phi_{in}^2 l}{2}}.$$

To account for the maximum degree d, we note that the uniform distribution π_{C_i} has l_2-norm $\sqrt{1/|C_i|}$, and the distribution p_v^l has total mass 1. The final bound becomes

$$\left\|p_v^l - \pi_{C_i}\right\|_2 \leq \sqrt{\frac{d}{|C_i|}} \cdot e^{-\frac{\phi_{in}^2 l}{2}}.$$

By setting $l = \Theta\left(\frac{\log n}{\phi_{in}^2}\right)$, we ensure that the distance decays to $\frac{\varepsilon}{4\sqrt{dn}}$, thereby guaranteeing sufficient mixing within each cluster. This bound implies that for any two vertices $u, v \in C_i$, their endpoint distribution satisfies

$$\left\|p_u^l - p_v^l\right\|_2 \leq \left\|p_u^l - \pi_{C_i}\right\|_2 + \left\|p_v^l - \pi_{C_i}\right\|_2 \leq \frac{\varepsilon}{2\sqrt{dn}}.$$

Such distance gap with two endpoints ensures that vertices within the same cluster are densely connected in the similarity graph H, while distinct clusters remain separated. Let $u, v \in C_i$, we have

$$\left\|p_u^l - p_v^l\right\|_2^2 \leq \frac{\varepsilon^2}{4dn}.$$

Now u and v are connected in the similarity graph H since

$$\left\|p_u^l - p_v^l\right\|_2^2 < \tau = \frac{\varepsilon^2}{16n} (d > 4).$$

Hence, vertices within the same cluster form a clique in H. Then suppose $u \in C_i$, $v \in C_j$, $i \neq j$. The distributions are far apart, and u and v are not connected in H by Lemma 3.4. Therefore, distinct clusters remain disconnected.

■

Lemma 3.3 (Construction of the similarity graph) Let S be a uniformly sampled vertex set, and let H be the similarity graph constructed following Steps 2 and 3 of the tester. If the input graph G is clusterable, then H can be partitioned into at most k connected components. Conversely, if G is ε-far from being clusterable, H will contain more than k connected components.

Proof An edge (u, v) is added to H if and only if the l_2^2-distance between their lazy random walk endpoint distributions p_u^l and p_v^l is below a threshold τ. The walk length l ensures sufficient mixing of

distributions, and the threshold is set as $\frac{\varepsilon^2}{16n}$. Under this configuration, $\left\|p_u^l - p_v^l\right\|_2 \leq \tau$ guarantees that vertices within the same cluster C_i form a clique in H, thereby ensuring H has at most k connected components. This lemma rigorously bridges the structural properties of G to the connectivity of H, forming the theoretical foundation for clusterability testing. In summary, the input graph G is $(k, \phi_{in},\ \phi_{out},\ \lambda_{in})$-clusterable if the similarity graph H is partitioned into at most k connected components with high probability $\frac{2}{3}$. ■

(Reject) The preceding analysis in Lemma 3.3 establishes that the similarity graph H reliably reflects the clusterability of G: if G is k-clusterable, H contains at most k connected components, whereas H contains more than k components when G is ε-far from being clusterable. To rigorously formalize the latter case, we must delve into the structural incoherence in ε-far instances. Specifically, the inability to partition G into k clusters with bounded conductance and label consistency implies the existence of $k+1$ or more disjoint subsets, each exhibiting weak internal-cluster connectivity and violating the label consistence constraints. Lemma 3.4 bridges this intuition to formal proof by constructing such subsets and quantifying their properties.

Crucially, the parameter choices, particularly the threshold τ, are designed to amplify the distributional divergence between vertices from distinct subsets. When G lacks a cohesive clustering structure, random walks originating from different subsets fail to mix, resulting in endpoint distributions with l_2^2-distance exceeding τ. This ensures that H cannot connect these subsets into fewer than $k+1$ components. Lemma 3.4 rigorously demonstrates this phenomenon, leveraging conductance bounds and sampling arguments to confirm the inevitability of fragmentation in H. This transition ensures the algorithm's robustness in distinguishing clusterable graphs from those requiring substantial modifications.

Lemma 3.4 If graph G is ε-far from being clusterable, there exists $k+1$ disjoint subsets $V_1, V_2, ..., V_k, V_{k+1}$ satisfying $\forall i, |V_i| \geq \frac{\varepsilon^2 n}{1152k}$ and $\phi_G(V_i) \leq O\left(\frac{\phi_{in}}{\varepsilon^2}\right)$, such that for any pair of vertices $u \in V_i$ and $v \in V_j$ $(i \neq j)$, when $l \leq \Theta\left(\frac{\log n}{\phi_{in}^2}\right)$, it holds

$$\left\|p_u^l - p_v^l\right\|_2^2 \geq \frac{1}{n}.$$

Proof Given the threshold $\tau = \frac{\varepsilon^2}{16n} < \frac{1}{n}$, vertices from distinct subsets V_i and V_j cannot be connected in H, as their endpoint distributions violate the l_2^2-distance criterion. Consequently, H must contain at least $k+1$ connected components, reflecting the structural fragmentation of G.

We use the partitioning Lemma [4]. If G is is ε-far from being clusterable, there exists a partition of V into $k+1$ subsets $V_1, ..., V_{k+1}$ such that $|V_i| \geq \frac{\varepsilon^2 n}{1152k}$. Now, we apply Coupon Collector Principle. We sample $s = \Theta\left(\frac{k \log n}{\varepsilon^2}\right)$ vertices. The probability that a particular subset V_i contains no sampled vertex is $\left(1 - \frac{|V_i|}{n}\right)^s \leq \exp\left(-s \cdot \frac{|V_i|}{n}\right) \leq \exp\left(-\frac{k \log n}{\varepsilon^2} \cdot \frac{\varepsilon^2}{1152k}\right) = n^{-1/1152}$. By union bound, the probability that every subset is missed is at most $(k+1) \cdot n^{-1/1152} \leq \frac{1}{6}$ for large n.

Next, we use Chernoff bound to ensure that the number of sampled vertices in each V_i is concentrated. Let X_i be the number of sampled vertices in V_i. Then expectation of X_i is

$$E[X_i] = s \cdot \frac{|V_i|}{n} \geq \frac{k \log n}{\varepsilon^2} \cdot \frac{\varepsilon^2}{1152k} = \frac{\log n}{1152}.$$

By Chernoff bound, we get

$$\Pr\left(X_i < \tfrac{1}{2} E[X_i]\right) \leq \exp\left(-\frac{E[X_i]}{8}\right) \leq n^{-1/9216}.$$

By union bound, the probability that all X_i are at least $\frac{\log n}{1152}$ is at least $1-(k+1)n^{-1/9216} \geq \frac{5}{6}$. Thus, the graph H will have at least $k+1$ connected components with probability at least $\frac{5}{6}$.

■

Theorem 3.5 (Label Sampling Error Control) Let $\widehat{C}_i$ be an estimated cluster in H and $\lambda_i \in \Sigma$ denote the dominant label for $\widehat{C}_i$. Suppose we uniformly sample $s' = \Theta\left(\frac{\log(k/\delta)}{\lambda_{in}^2 \varepsilon^2}\right)$ labels (edges) from $\widehat{C}_i$. Let $\hat{\lambda}_i$ be the empirical proportion of the dominant label in the sampled edges. Then, with probability at least $1-\delta$, the estimation error satisfies

$$\forall i, \left|\widehat{\lambda}_i - \lambda_i\right| < \frac{\varepsilon}{2},$$

where λ_i is the true dominant label proportion in $\widehat{C}_i$.

Proof Let us fix one edge $e \in \widehat{C}_i$, we define an indicator variable Y_j such that if the label of the j-th sampled edge equals σ_i, $Y_j = 1$, else $Y_j = 0$.

The true dominant label proportion is $\lambda_i = E[Y_j]$, and the empirical estimate is $\widehat{\lambda}_i = \frac{1}{s'} \sum_{j=1}^{s'} Y_j$. Since $Y_j \in \{0, 1\}$ are independent and identically distributed random variables, Chernoff bound is applied to guarantee

$$\begin{aligned} & \Pr\left[\left|\widehat{\lambda}_i - \lambda_i\right| \geq \frac{\varepsilon}{2}\right] \\ = \; & \Pr\left[\left|\frac{1}{s'}\sum_{j=1}^{s'} Y_j - \lambda_i\right| \geq \frac{\varepsilon}{2}\right] \\ \leq \; & 2\exp\left(-\frac{s'\varepsilon^2}{2}\right). \end{aligned}$$

To bound this probability by δ, we just need to set

$$s' \geq \frac{2\ln(2/\delta)}{\varepsilon^2}$$

so that $2\exp\left(-\frac{s'\varepsilon^2}{2}\right) \leq \delta$.

Incorporating the dependency on λ_{in} (to conservatively account for rare labels), the required sample size becomes $s' = \Theta\left(\frac{\log(k/\delta)}{\lambda_{in}^2\varepsilon^2}\right)$.

■

3.3 Joint Error Probability Analysis

To ensure the overall reliability of the algorithm, we analyze the joint error probability arising from both connectivity estimation and label consistency verification. Theorem 3.6 is our main theorem.

Theorem 3.6 Given a labeled graph $G = (V, E, L)$ and a set of parameters (k, ϕ_{in}, ϕ_{out}, λ_{in}, ε). If graph G is (k, ϕ_{in}, ϕ_{out}, λ_{in})-clusterable, the tester(i.e., Algorithm 1) accepts with probability at least $\frac{2}{3}$. If G is ε-far from being clusterable, the tester rejects with probability at least $\frac{2}{3}$.

Proof Let E_1 denote the event of connectivity estimation error (i.e., incorrectly accepting or rejecting the graph based on structural properties), and E_2 denote the event of label consistency estimation error (i.e., misjudging the dominant label proportion in any cluster).

By Lemma 3.1, the probability of E_1 is bounded as

$$\Pr[E_1] \leq \tfrac{1}{6}.$$

This follows from Chernoff bounds on vertex sampling and the mixing time guarantee of lazy random walks.

By Theorem 3.5, for each cluster $\widehat{C}_i$, the probability of misestimating λ_i by more than $\frac{\varepsilon}{2}$ is bounded by $\delta = \frac{1}{6k}$. Applying the union bound over all k clusters, it holds

$$\Pr[E_2] = \Pr\left[\exists i, \left|\widehat{\lambda}_i - \lambda_i\right| \geq \tfrac{\varepsilon}{2}\right] \leq k\cdot\delta = k\cdot\tfrac{1}{6k} = \tfrac{1}{6}.$$

By setting $\delta = \frac{1}{6k}$, the total error probability is bounded by $\frac{1}{6}$. In conclusion, the algorithm terminates successfully with probability

$$\begin{aligned} & \Pr[\text{The tester accepts } G] \\ \geq \; & 1 - (\Pr[E_1] + \Pr[E_2]) \\ = \; & 1 - \left(\frac{1}{6} + \frac{1}{6}\right) = 1 - \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

Similarly, the tester rejects G which is ε-far from being clusterable with probability

$$\begin{aligned} & \Pr[\text{The tester rejects } G] \\ \geq \; & \Pr\left[E_2|\overline{E_1}\right] \\ \geq \; & \frac{2}{3}. \end{aligned}$$

■

3.4 Time Complexity and Query Complexity

Now we analyze the running time complexity of algorithm Label-aware Clusterability Tester. Algorithm 1 consists of two core phases: connectivity estimation and label consistency verification. For each sampled vertex s, m lazy random walks traverses l steps, yielding total walk time of $O(s\cdot m\cdot l) = O\left(\frac{k\log^3 n}{\varepsilon^2\phi_{in}^2}\right)$. Pairwise l_2^2-distance comparisons between s vertices involve $O(s^2)$ pairs. Each comparison computes distances between m-dimensional distributions ($O(m)$ per pair), resulting in time $O\left(s^2\cdot m\right) = O\left(\frac{k^2\log^3 n}{\varepsilon^4}\right)$. The empirical label ratio per cluster requires $O(k\cdot s')$ time, dominated by sampling. Combining all phases, the total time complexity is $O(\frac{k\log^3 n}{\varepsilon^2\phi_{in}^2} + \frac{k^2\log^3 n}{\varepsilon^4} + \frac{k\log k}{\lambda_{in}^2\varepsilon^2}) = \tilde{O}(\sqrt{n}\cdot\mathrm{poly}(k, 1/\phi, 1/\varepsilon))$. Since the time of the sampling steps is constant, the query complexity is also $\tilde{O}(\sqrt{n})$.

4 Summary and Future Work

This paper presents a novel framework for testing clusterability on labeled graphs, addressing the critical limitation of existing methods that ignore semantic information from edge labels. We introduce a label-aware clusterability criterion requiring clusters to satisfy both structural conductance bounds and labels consistency, enabling interpretable clustering in applications like social networks and bioinformatics. The sublinear-time algorithm integrates label-augmented random walks and estimation strategies for connectivity and label distribution verification. This work bridges the gap between traditional graph clustering and the emerging field of labeled graph analysis, providing a theoretical foundation for testing clusterability in labeled networks. Future work will

extend this framework to dynamic graphs with adaptive sampling for evolving labels and structures, generalize it to heterogeneous graphs with multi-type labels (e.g., temporal or hierarchical), and investigate tighter lower bounds under label-topology correlations. Integrating our method with graph neural networks may improve hierarchical pooling or explainability by aligning latent representations with semantic clusters.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62272280 and Grant 61972228.

Declaration of Interest

The authors of this work declare no conflict of interest.

References

[1] Gajula Ramesh, Anil Kumar Budati, Shayla Islam, Louai A. Maghrabi and Abdullah Al-Atwai, Artificial intelligence enabled future wireless electric vehicles with multi-model learning and decision making models, *Tsinghua Science and Technology,* vol. 29, no. 6, pp. 1776–1784, 2024.

[2] Goldreich Oded and Ron Dana, A sublinear bipartiteness tester for bounded degree graphs, in *Proceedings of the thirtieth annual ACM Symposium on Theory of Computing*, 1998, pp. 289–298.

[3] Goldreich Oded and Ron Dana, Property testing in bounded degree graphs, in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, 1997, pp. 406–415.

[4] Artur Czumaj, Pan Peng and Christian Sohler, Testing cluster structure of graphs, in *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, 2015, pp. 723–732.

[5] Artur Czumaj, Pan Peng and Christian Sohler, Relating two property testing models for bounded degree directed graphs, in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016, pp. 1033–1045.

[6] Noga Alon and Asaf Shapira, Every monotone graph property is testable, in *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 2005, pp. 128–137.

[7] Junfen Chen, Jie Han, Xiangjie Meng, Yan Li and Haifeng Li, Graph convolutional network combined with semantic feature guidance for deep clustering, *Tsinghua Science and Technology,* vol. 27, no. 5, pp. 855–868, 2022.

[8] Chaoqi Jia, Longkun Guo, Kewen Liao and Zhigang Lu, Efficient algorithm for the k-means problem with must-link and cannot-link constraints, *Tsinghua Science and Technology,* vol. 28, no. 6, pp. 1050–1062, 2023.

[9] Chiplunkar Ashish, Kapralov Michael, Khanna Sanjeev, Mousavifar Aida and Peres Yuval, Testing graph clusterability: Algorithms and lower bounds, in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018, pp. 497–508.

[10] Gluch Grzegorz, Kapralov Michael, Lattanzi Silvio, Mousavifar Aida and Sohler Christian, Spectral clustering oracles in sublinear time, in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2021, pp. 1598–1617.

[11] Pan Peng and Yuyang Wang, An optimal separation between two property testing models for bounded degree directed graphs, https://arxiv.org/abs/2305.13089, 2023.

[12] Forster Sebastian, Nanongkai Danupon, Yang Liu, Saranurak Thatchaphol and Yingchareonthawornchai Sorrachai, Computing and testing small connectivity in near-linear time and queries via fast local cut algorithms, in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2020, pp. 2046–2065.

[13] Kumar Akash, Seshadhri C and Stolman Andrew, Random walks and forbidden minors II: a poly(dε-1)-query tester for minor-closed properties of bounded degree graphs, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019, pp. 559–567.

[14] Yifei Li, Donghua Yang and Jianzhong Li, Testing higher-order clusterability on graphs, *Journal of Combinatorial Optimization,* vol. 49, no. 3, pp. 1–21, 2025.

Pengzhi Gao is a PhD candidate at the School of Software, Shandong University, under the supervision of Professor Peng Zhang. He is interested in graph algorithms, sublinear algorithms and related problems of property testing.

Peng Zhang received the B. Eng. and M.S. degrees in computer science from Shandong University, Jinan, China, in 1999 and 2004, respectively, and the Ph.D. degree in computer science from the Institute of Software, Chinese Academy of Sciences, in 2007. He is currently a Professor with School of Software, Shandong University, Jinan, China. His research interests include algorithm design and analysis, combinatorial optimization, and computational complexity.