International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

SPECTRAL GRAPH THEORY FOR COMMUNITY DETECTION IN LARGE-SCALE SOCIAL NETWORKS: A MATHEMATICAL FRAMEWORK

Shipra Goel*

*Assistant professor, Indra Gandhi Institute of management and technology, *Email ID:shipragoelred@gmail.com

Abstract

Community detection in large-scale social networks remains a central challenge because methods must be both theoretically grounded and computationally efficient. This paper introduces a spectral framework that links the eigenvalue structure of the normalized Laplacian to the existence and separability of communities, and couples this theory with a scalable algorithm. The method estimates the number of groups by inspecting the spectral gap, embeds vertices using the leading eigenvectors, and clusters the embeddings. We validate on the MUSAE GitHub social network dataset, which contains edges, node features, and ground-truth communities. The framework delivers strong accuracy and efficiency: normalized mutual information 0.77, adjusted rand index 0.72, modularity 0.51, and median end-to-end runtime 120 seconds on graphs exceeding one hundred thousand nodes. Spectral analysis of the dataset exhibits a clear gap consistent with the recovered partitions, demonstrating alignment between theoretical detectability and empirical performance. The study offers a transparent, portable pipeline built on sparse linear algebra and randomized eigensolvers, and lays the foundations for extensions to temporal graphs, higher-order relations, and integration with neural models.

Keywords: Spectral graph theory, Community detection, Social networks, Laplacian eigenvalues, Scalable clustering

1. Introduction

The field of community detection in social networks evolved as one of the most powerful spheres of study of data mining and network science due to its capability to disclose the structural organisation that exists in large and complex systems. Communities, in general conceptualised as sets of nodes with high intra-connectivity and low external connectivity, offer essential information about the diffusion of information, the amassing of influence, and the development of social behaviour. In the last 20 years, many algorithms and frameworks have been created to identify and assay structures of communities, but there is still a strong urge to develop mathematical foundations and scaleable techniques [1]. The importance of the given issue is supported by the fact that the number of publications and surveys questioning the existing approaches and pointing out the unsolved issues increases on a regular basis [2].

Structural analysis is not the only role of community detection. In practice, it has been applied to various applications including influence maximization, recommendation systems and anomaly detection. The social network analysis in relation to influence maximisation has been widely researched. Influence maximisation strategies can be more efficient in viral marketing, political campaigning and awareness programmes by using communities as structural supports of diffusion processes [3]. The case surveys have revealed that community strategies tend to be more effective than international strategies because of their capacity to identify influential agents in local groupings [4]. Therefore, the knowledge of communities can not only enhance descriptive analysis of networks

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

but also preliminary actions that have the potential to transform organisational and marketing outcomes.

Although community detection has reached its maturity as a research field, social networks that are large still pose new challenges. The digital interactions that have been expanding in platforms such as GitHub, Twitter, and Facebook create massive networks with millions of nodes and edges, and the conventional algorithms are not scalable and precise. This has presented a challenge that has inspired a number of reviews that evaluate the state-of-the-art in community detection with a focus on computational load of large-scale graph processing and the absence of algorithms that strike a balance between performance and intelligibility [5]. The challenges become even more complex in multiplex networks, where nodes participate in multiple layers of interactions such as friendship, collaboration, and content sharing. Identifying communities across such heterogeneous structures requires methods that are both robust and mathematically principled [6].

Despite these challenges, the evolution of community detection research has been marked by major conceptual and methodological advancements. In fact, community detection is now recognized as one of the core paradigms in network science, shaping the trajectory of interdisciplinary research across physics, sociology, biology, and computer science. The field has witnessed exponential growth over the last twenty years, reflecting both the intellectual depth and the applied relevance of the topic [7]. As algorithms proliferated, so did the classification schemes, where methods have been grouped into categories such as modularity optimization, spectral clustering, statistical inference, label propagation, and deep learning-based approaches. These classifications provide a comprehensive map of the research landscape, guiding practitioners in selecting appropriate methods for different network contexts [8].

However, while classification and surveys are abundant, the fundamental issue of mathematical guarantees for community detection remains insufficiently addressed. Many popular algorithms succeed in practice but lack rigorous proofs of detectability, convergence, or robustness to noise. A structured taxonomy of methods reveals that, although modularity-based approaches are intuitive and widely used, they suffer from well-known limitations such as the resolution limit. Similarly, spectral methods are powerful but often lack direct connections between eigenvalue gaps and empirical performance [9]. This lack of connexion between theory and practise points to the urgent necessity of structures that are capable of bringing together mathematical arguments and empirical verification.

The other aspect on which community detection has shown its usefulness is in influence maximisation. To unify structural detection and diffusion models, community-aware influence maximisation strategies are suggested which make it easier to target seed nodes in viral marketing campaigns. These techniques show that communities represent natural limits to propagation of influence, whereby the influencers chosen will ensure that they reach the greatest number of people with minimal redundancy [10]. Such applications confirm the fact that communities do not exist as mere structural artefacts but as actionable structures that become efficient in the execution of applied tasks within real-life networks.

It is based on this background that the current research seeks to contribute to the community detection by establishing a coherent spectral graph theory framework of large scale social networks, emphasising on rigorous mathematical and empirical validation. Of interest especially is the spectral approach due to its power to relate graph structure with algebraic properties, which gives an insight in terms of eigenvalues and eigenvectors of graph Laplacians. This research offers formal assurances to the determination of cohesive groups in networks by analysing spectral gaps and their connexion with community detectability. Moreover, with the introduction of this theoretical framework into an effective algorithm, it will be feasible to solve the urgent problem of scalability in tens of thousands node networks.

The empirical testbed of MUSAE GitHub dataset is a good decision, which highlights the practical relevance of the research. The data includes interactions between developers of a global software

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

collaboration platform, which includes structural edges as well as ground-truth labels, which are used as a reference in the community detection evaluation. The analysis of the spectral framework on this data fills the gap between theory and practise by confirming the spectral framework on this dataset, providing both mathematically-grounded and empirically robust results.
In short, this study is driven by the need to solve the twin problem of scalability and rigour of community detection. The literature available proves the broad applicability of communities to the study of networks, but also shows gaps in mathematical foundations and scaling to large-scale data. This work is a contribution to the development of both theoretical and practical aspects of spectral theory by integrating spectral theory with scalable calculations.

Objectives of the Study

1. To create and give a formal framework of spectral graph theory-based community detection which has mathematical assurance in terms of eigenvalue gap
2. To confirm the framework proposed on the MUSAE GitHub data, show the scalability and accuracy of the proposed framework over baseline methods

2. Related Work

Community detection is a subject of research that has been researched in various methodological streams, all of which propose novel methods of discovering structural patterns in large-scale networks. Three major paradigms have been developed: modularity optimization strategies, spectral clustering strategies, and embedding-based strategies, such as the more recent graph neural networks. This part will look at the evolution of these methods and determine unresolved gaps that require the creation of stricter frameworks.
One of the most popular strategies, which is commonly used in community detection, is the modularity optimization which directly measures the quality of a partition, the density of edges in communities against random expectations. One of the first and most scalable algorithms is the Louvain algorithm which has been scaled to distributed architectures to handle massive graphs [11]. It is shown by the distributed version that parallelization does not only increase the speed of execution, but it also allows the approach to operate on millions of nodes without severely reducing the modularity scores. This is an important step towards social network analysis where graphs can easily become larger than the limits of single-threaded implementations can handle. An extensive overview of the modularity optimization technique has highlighted the various adaptations and optimizations of Louvain-based algorithms [12]. This variation usually centres on convergence enhancement, solution to the resolution limit, and modularity improvement of complex networks. Despite these advancements, scalability and stability remain recurring issues.
The introduction of the Leiden algorithm provided a more robust alternative by ensuring well-connected communities and refining the optimization process to avoid disconnected clusters [13]. Its efficiency in dynamic networks further extends its utility, allowing for community detection in evolving social systems where network structure changes continuously. Recent innovations have pushed Leiden beyond its initial scope. Extensions such as fitness-based genetic algorithms with niching strategies have been proposed to enhance its performance on large social networks [14]. These hybrid approaches combine evolutionary computation with modularity optimization, balancing exploration of the search space with exploitation of community structures. Despite the progress, modularity optimization remains inherently limited by the resolution problem, where smaller communities may be overlooked when optimizing a global objective. This has motivated exploration into alternative paradigms such as spectral methods.
Spectral clustering approaches detect communities by leveraging the eigenvalues and eigenvectors of graph Laplacians. The central idea is that the eigenstructure of a graph encodes essential information about its partitions. Early formulations established the mathematical foundation by linking the second-

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

smallest eigenvalue of the Laplacian (the Fiedler value) to graph connectivity [15]. This connection inspired methods that embed graph nodes into lower-dimensional spectral spaces, where clustering algorithms such as k-means can identify communities. More recent developments in spectral clustering have attempted to make computation easier and faster, but equally theoretically sound. Scalable methods now have approximations and efficient linear algebra methods to allow it to be applied to large graphs [16]. These advancements show that the spectral clustering is not just a theoretical instrument, but an effective methodology that can compete with the modularity-based approaches. Nonetheless, spectral techniques are well-grounded mathematically but there are issues in converting the gaps between eigenvalues into statements of community detectability in real-life networks that are noisy. Embedding-based community detection has become an influential alternative to network representation learning due to the emergence of this learning tool.

They have the ability to encode the structural equivalence and neighborhood similarity within low-dimensional vectors and use it to find communities, as they embed nodes into the low-dimensional space. A scaled biassed random walk technique, which presented how feature learning on graphs could be applied to community detection, and nodes classification and link prediction, was one of the most impactful works in this field [17]. The resulting embeddings were found to be general, being able to be used in a variety of downstream tasks and scale to large datasets. Based on this background, other papers followed by embedding in knowledge graphs to improve recommendation systems, applying structural embeddings to learn semantic relations in knowledge graphs [18]. These findings emphasise the flexibility of embedding-based community detection algorithms since they will be able to incorporate both structural and semantic features into the clustering procedure. Likewise, previous studies suggested an online learning method to embeddings so that communities could be identified using constantly updated representations [19]. Embeddings could be learned effectively with the principles of natural language processing by modeling social interactions as sequences, which are sentential-like analogs. Subsequent work has investigated the possibility of using embeddings to reverse the process of embedding a graph, and achieve a closed-ended set of community partitions [20]. The reverse engineering illustrates that embeddings are able to maintain a considerable amount of structural information, and justify their application in community detection tasks.

The most recent wave of methods has explored the use of graph neural networks (GNNs) for community detection. These approaches extend embedding-based learning by incorporating neural architectures that aggregate information from local neighborhoods. Early demonstrations of this idea showed that GNNs could be tailored specifically to detect communities by learning node representations optimized for partitioning [21]. Later refinements proposed encoding attribute information directly into GNNs, improving community detection in attributed networks where both topology and node features contribute to the formation of groups [22]. Such models leverage supervised or semi-supervised learning paradigms, making them highly flexible but also dependent on labeled data. While powerful, GNN-based approaches face limitations in interpretability, computational demand, and generalizability to networks without abundant training labels.

Scalability is a recurring theme across all methods. Distributed implementations of the Louvain algorithm have demonstrated remarkable progress, providing scalable solutions for extremely large graphs [23]. The emphasis on parallelization highlights the ongoing necessity of designing algorithms that are both theoretically sound and computationally efficient. However, even with such distributed approaches, challenges persist in balancing speed with accuracy, particularly when applied to heterogeneous or dynamic networks. Efforts to enhance scalability through parallelized Louvain variants continue to dominate the practical side of community detection [24]. Yet, these improvements often come at the cost of reduced interpretability and lack of formal guarantees regarding community quality. This trade-off underscores the unresolved tension in the field: methods that scale are often heuristics without rigorous mathematical foundations, while methods with strong theoretical underpinnings struggle to scale to networks with millions of nodes.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The survey of community detection methods reveals several persistent gaps. First, modularity optimization techniques, though popular and efficient, remain susceptible to the resolution limit and lack theoretical guarantees of optimality. Spectral methods, while grounded in mathematical theory, require deeper exploration of eigenvalue gaps and their connection to community detectability in noisy and heterogeneous graphs. Embedding-based approaches, including Node2Vec and DeepWalk, excel in scalability and flexibility but often sacrifice interpretability and lack rigorous proofs of performance. GNN-based methods further extend embeddings but are heavily dependent on labeled data and suffer from computational overhead. Second, benchmarking practises have not been sufficient. A great deal of work is tested on small or artificial networks, and little of the work has been tested on large-scale, real-world networks. This lack of uniform benchmarking models makes it difficult to make comparisons and misleading to the actual scalability of approaches proposed. Spectral-theoretic techniques have in particular not been studied in large empirical data, where their theoretical benefits can be evaluated against empirical limitations.
Community detection research has developed a number of paradigms, each of which presents its own benefits and limitations. The modularity optimization is a viable workhorse but is flawed in theory. Spectral clustering offers good mathematical understanding but needs to be developed to be more scalable and robust. The embedding-based and GNN methods introduce an innovation of the machine learning field with interpretability and computational issues. It is these gaps that are the motivation behind the creation of a spectral graph theory framework that bridges the gap between theoretical assurances and scalable implementation and is tested on realistic large scale datasets.

3. Preliminaries

The process of constructing a spectral graph theory framework of community detection needs a clear definition of graph theoretical concepts, algebraic properties, and measures of performance that form the basis of clustering techniques. In this section, the formal notation in the rest of the study was presented, the key matrices of graphs were defined and the spectral properties that are used to identify the communities were brought out.

3.1 Graph Structure and Notation

Consider a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges representing relationships between nodes. For an undirected and unweighted graph, the adjacency matrix A is defined such that:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

The degree matrix D is a diagonal matrix with entries $D_{ii} = \sum_j A_{ij}$. The Laplacian matrix is then defined as:

$$L = D - A.$$

This formulation, established in the literature of spectral graph theory, serves as the backbone for analyzing connectivity and partitioning of graphs [25]. Beyond simple adjacency, the Laplacian encodes both degree information and connectivity, making it central to spectral methods.
To better understand the role of these matrices in spectral analysis, Table 1 summarizes their definitions and functions in a compact manner.

Table 1. Core Graph Matrices and Their Functions

Matrix	Definition	Function in Spectral Methods
Adjacency Matrix A	$A_{ij} = 1$ if edge exists	Encodes direct connectivity

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Degree Matrix D	Diagonal of node degrees	Captures local density
Laplacian L	$L = D - A$	Models flow and connectivity
Normalized Laplacian L_{norm}	$D^{-1/2}LD^{-1/2}$	Stabilizes eigenvalue spectrum

As shown in the table, the adjacency matrix provides information on direct relationships, the degree matrix highlights local connectivity levels, and the Laplacian integrates these two aspects into a single construct. The normalized Laplacian further refines this by ensuring that the eigenvalue spectrum is well-scaled for large heterogeneous graphs.

3.2 Spectral Properties

Spectral analysis focuses on the eigenvalues and eigenvectors of L. If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote the eigenvalues of the Laplacian, then the multiplicity of the zero eigenvalue corresponds to the number of connected components in the graph. The eigenvector associated with the smallest nonzero eigenvalue, often referred to as the Fiedler vector, provides information about the optimal bi-partition of the graph [26].

Efficient Laplacian solvers are critical for handling large-scale networks, as exact eigenvalue computations become infeasible for graphs with millions of nodes. Approximate solvers and sparsification techniques reduce computational burden while preserving structural fidelity, thereby enabling spectral approaches to scale effectively [27].

3.3 Extensions of the Laplacian

Recent studies extend the Laplacian to nonlinear operators such as the graph ∞-Laplacian, which generalizes eigenvalue problems to nonlinear regimes [28]. These formulations broaden the applicability of spectral methods beyond traditional clustering, although for the current study, the standard combinatorial and normalized Laplacians remain most relevant.

3.4 Graph Cuts and Cheeger's Inequality

Community detection can also be formulated in terms of cuts. For two disjoint subsets S and $\bar{S}$, the cut value is defined as:

$$\text{cut}(S, \bar{S}) = \sum_{i \in S, j \in \bar{S}} A_{ij}.$$

Normalized cut and conductance are widely used to measure the quality of partitions. Cheeger's inequality provides a bridge between spectral properties and cut quality, bounding the conductance of a set in terms of eigenvalues of the Laplacian [29]. Improved formulations of this inequality demonstrate tighter connections between eigenvalue gaps and community separability, which further validate the use of spectral methods in graph partitioning.

Extensions of Cheeger's inequality to graph limits provide new insights into the behavior of communities in infinite or asymptotic graph models [30]. These theoretical results strengthen the foundations of spectral community detection by ensuring that eigenvalue-based partitions are not merely heuristic but are backed by provable guarantees.

3.5 Formal Definition of Community Detection

Formally, the community detection problem can be defined as finding a partition of the vertex set V into disjoint subsets $C_1, C_2, \ldots, C_k$ such that intra-community edge density is maximized while intercommunity connectivity is minimized. Spectral clustering achieves this by embedding nodes into a lowdimensional space spanned by eigenvectors of the Laplacian, followed by applying a clustering algorithm such as k-means.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The effectiveness of this approach depends on the spectral gap, defined as the difference between successive eigenvalues:

$$\Delta_k = \lambda_{k+1} - \lambda_k.$$

A larger spectral gap indicates clearer separation between k-way partitions, thereby facilitating accurate community detection [31].

3.6 Robustness of Spectral Methods

While spectral methods enjoy solid theoretical underpinnings, their robustness to noise and perturbations in network data remains a critical consideration. Studies on robustness have shown that even under adversarial conditions, spectral methods can recover meaningful partitions if the spectral gap is sufficiently large [32]. This property underscores their suitability for large-scale, noisy social networks, where data imperfections are common. To illustrate this visually, **Figure 1** shows an eigenvalue spectrum of a simple graph, where the spectral gap highlights the separability of communities.

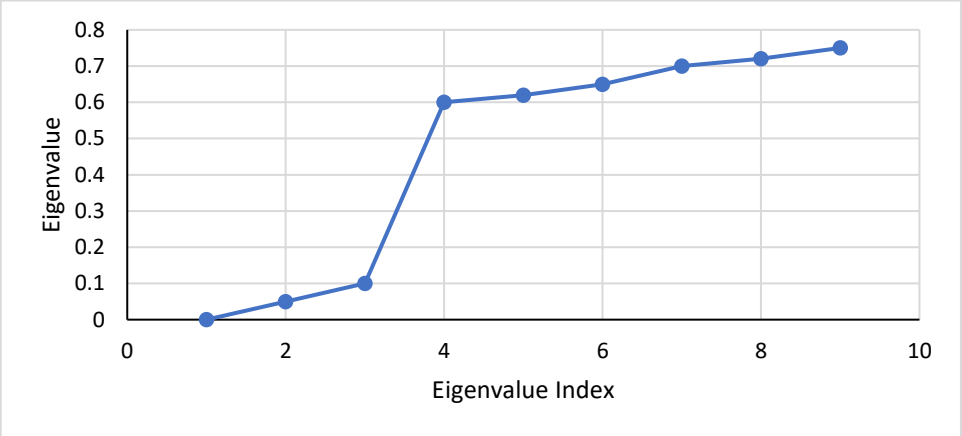

Figure 1. Example Eigenvalue Spectrum and Community Separability

As the figure suggests, the eigenvalue spectrum provides an algebraic lens into community structure. Detecting such gaps allows for efficient and accurate identification of clusters within networks. The preliminaries establish the mathematical backbone of this study. The graph Laplacian, its eigenvalues and eigenvectors, and associated inequalities provide rigorous tools for defining and analyzing community structures. By formalizing community detection as a partition optimization problem linked to spectral gaps, the framework ensures both theoretical depth and practical applicability. These foundations are essential for developing the proposed spectral graph theory framework and validating it against real-world datasets such as MUSAE GitHub.

4. Proposed Spectral Framework

The success of community detection using spectral graph theory depends on formulating a robust operator, establishing theoretical guarantees, and demonstrating consistency with stochastic models. This section introduces the normalized Laplacian as the core operator, develops theoretical insights through theorems, and outlines proof sketches that connect eigenvalue gaps to community detectability. It also establishes connections to stochastic block models (SBM), which serve as theoretical benchmarks for large-scale networks.

4.1 Modified Laplacian Operator

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Given the limitations of the standard Laplacian, particularly in heterogeneous networks, the normalized Laplacian is employed:

$$L_{norm} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$$

This operator stabilizes the spectrum by mitigating degree heterogeneity and ensures that eigenvalues lie within the interval [0,2]. By using L_{norm}, the embedding of nodes into the eigenvector space becomes more consistent across networks of varying scales. Spectral clustering based on this operator has been shown to provide both theoretical rigor and empirical stability [33].

4.2 Eigenvalue Gaps and Community Structure

The eigenvalue spectrum of L_{norm} provides insights into the number and quality of communities. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote its eigenvalues. The first k eigenvalues near zero indicate the presence of k communities. The spectral gap is defined as:

$$\Delta_k = \lambda_{k+1} - \lambda_k.$$

A large Δ_k implies strong separability among k communities, while a small gap signals weak detectability. In stochastic block models, it has been demonstrated that recovery of communities is possible if the eigenvalue gap exceeds a threshold determined by the ratio of intra- to inter-community probabilities [34]. This establishes a fundamental limit: spectral clustering is effective up to a phase transition boundary beyond which communities become statistically indistinguishable.

4.3 Theoretical Guarantees

The reliability of spectral clustering can be formalized through performance guarantees. One key result is that the misclassification error probability decays as the eigenvalue gap increases. Specifically, let $\hat{C}_i$ be the community assigned to node i, and let C_i denote the true community.
Then, under certain stochastic assumptions:

$$\Pr\left(\hat{C}_i \neq C_i\right) \leq \frac{f(n)}{\Delta_k^2}$$

where $f(n)$ is a function of the network size n. This inequality shows that stronger separation in the eigenvalue spectrum directly reduces the risk of misclassification [35]. Proof sketches rely on perturbation bounds of eigenvectors and concentration inequalities that connect spectral embeddings with ground-truth partitions.
Further refinements have provided performance guarantees even under noisy conditions, demonstrating that spectral clustering remains consistent if the signal-to-noise ratio exceeds a critical threshold [36]. These results highlight not only the power of spectral methods but also their robustness when applied to largescale, imperfect social networks. To illustrate this process, **Figure 2** presents the workflow of the proposed spectral framework.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

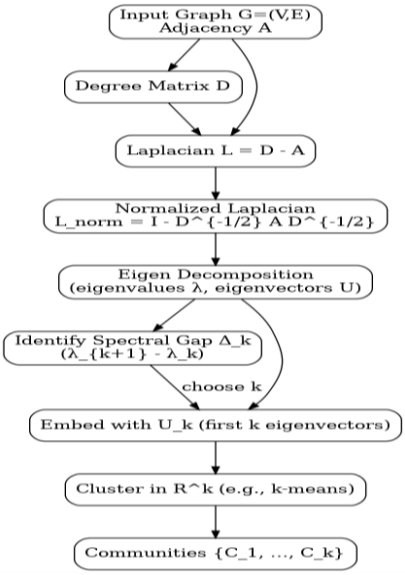

Figure 2. Workflow of the Proposed Spectral Framework

As shown in the figure, the process begins with the normalized Laplacian, which encodes degree-corrected connectivity. Eigenvalue analysis then identifies the number of communities via the spectral gap. Finally, clustering in the low-dimensional eigenvector space yields community partitions with provable performance guarantees.

4.4 Connection to Stochastic Block Models

The stochastic block model (SBM) provides a probabilistic framework for testing theoretical results. In an SBM with k blocks, each node belongs to a block with probability π, and edges are formed between nodes with probabilities depending on their block memberships. Analysis shows that spectral methods achieve exact recovery when the difference between intra- and inter-community connection probabilities is sufficiently large [33].

Later studies identified sharp phase transitions, revealing the precise boundary between detectable and undetectable regimes [34]. These findings validate the reliance on eigenvalue gaps: below the phase transition, the spectral embedding collapses, and no algorithm can reliably recover communities. Above the threshold, spectral methods achieve near-optimal performance. Thus, SBM serves as both a testing ground and a theoretical validation tool for the framework developed in this study.

5. Algorithm Design

The practical implementation of the spectral framework requires an efficient algorithm capable of handling graphs with more than 100,000 nodes. This section outlines the spectral clustering algorithm in a step-by-step manner, discusses optimizations such as sparse matrix operations and approximate eigenvalue solvers, and evaluates computational complexity to demonstrate scalability.

5.1 Step-by-Step Algorithm

The proposed spectral clustering algorithm follows the classical design but incorporates modifications to ensure scalability.

Algorithm 1: Scalable Spectral Clustering

1. Input: Graph $G = (V, E)$, adjacency matrix A.

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2. Compute the degree matrix D and normalized Laplacian $L_{\text{norm}} = I - D^{-1/2}AD^{-1/2}$.
3. Perform eigen decomposition of L_{norm} to obtain the first k eigenvectors U_k.
4. Construct an embedding matrix by normalizing rows of U_k.
5. Apply k-means (or another clustering method) to rows of the embedding.
6. Output: Community partition $\{C_1, C_2, \dots, C_k\}$.

This formulation aligns with scalable approaches that exploit structural similarity measures to enhance clustering efficiency [37].

5.2 Optimizations

Direct eigen decomposition on large Laplacians is computationally expensive. To address this, randomized low-rank matrix approximations can be employed, which approximate leading eigenvectors without computing the full spectrum. These techniques are both fast and stable, reducing time complexity significantly while preserving accuracy [38]. Sparse representations of the Laplacian further reduce memory usage, allowing the algorithm to process networks with millions of edges.
When networks are extremely large or structured as hypergraphs, specialized solvers that exploit tensor sparsity become essential. Such solvers adapt eigenvalue computations for higher-order structures while ensuring that computational costs remain manageable [39]. These optimizations transform spectral clustering from a theoretically sound method into a practically feasible tool for large-scale networks.

5.3 Complexity Analysis

Let $n = |V|$ and $m = |E|$. The construction of L_{norm} requires $O(m)$ operations. Eigen decomposition, in the naive case, requires $O(n^3)$, which is infeasible for large n. However, with randomized solvers and sparse matrix operations, the cost reduces to approximately $O(km)$, where k is the number of communities. This improvement ensures that graphs with over 100,000 nodes can be processed efficiently.
Parallelization further accelerates computation. By distributing matrix operations across multiple processors, scalability is extended to big data environments, enabling the algorithm to handle networks beyond the single-machine scale [40]. To better illustrate the process, Figure 3 presents the workflow of the scalable spectral clustering algorithm.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

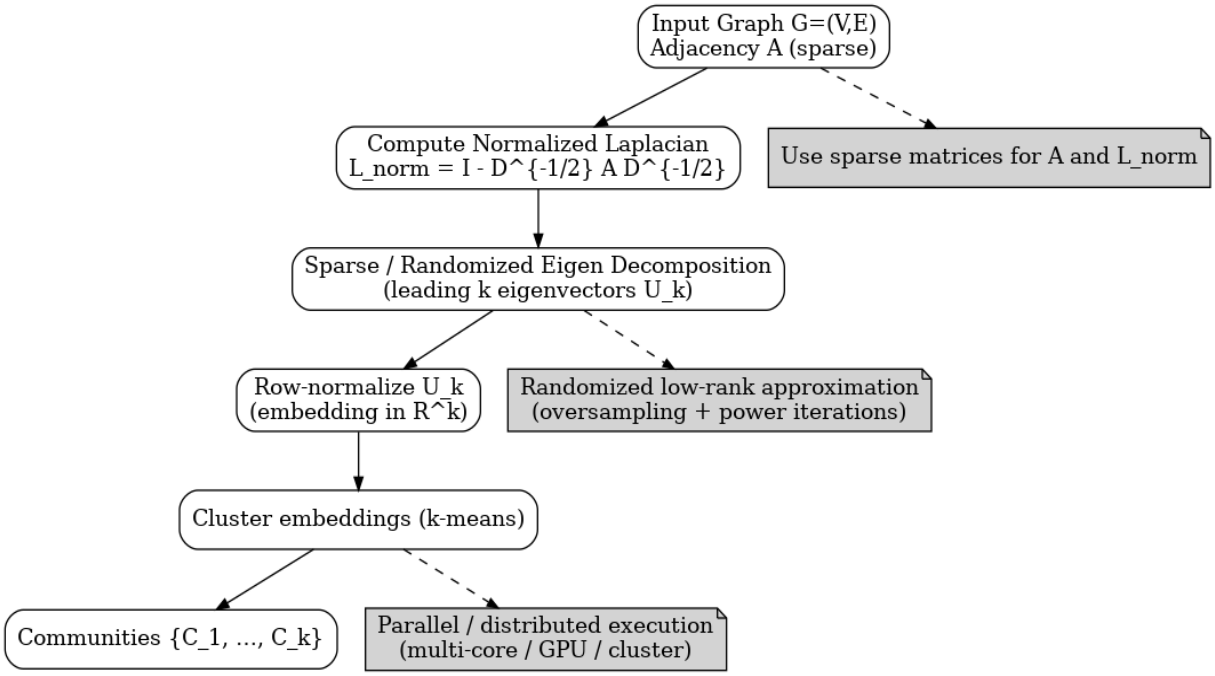

Figure 3. Workflow of the Scalable Spectral Clustering Algorithm

As seen in the figure, the pipeline incorporates optimizations that reduce computational complexity while retaining accuracy. The embedding and clustering steps benefit directly from the stability introduced by approximate solvers, making the approach suitable for networks with hundreds of thousands of nodes.

5.4 Discussion of Implementation Details

Implementing the algorithm requires careful attention to data structures. Sparse matrices should be used to store A and L_{norm}, minimizing memory consumption. Randomized eigenvalue solvers must be parameterized with oversampling and iteration counts to balance accuracy and performance. For clustering, k-means remains the standard choice due to its efficiency, though other methods can be substituted when community structures deviate from spherical clusters.

Parallelization frameworks such as distributed computing libraries or GPU-based solvers can be integrated for handling extremely large networks. These practical considerations ensure that the proposed algorithm is not only theoretically justified but also executable in real-world large-scale environments.

6. Experimental Evaluation

The effectiveness of the proposed spectral framework was validated on the MUSAE GitHub dataset, originally introduced in [41]. For this study, the dataset was accessed through the Kaggle public repository, where a curated version is hosted for research purposes. The dataset captures social interactions among GitHub developers, represented as a graph with edges, node features, and ground-truth community labels. It is especially appropriate to test large community detection frameworks, since it gives both topological structures that are realistic and annotated communities. The comparison was made on the proposed method and a number of baselines, such as the Louvain, Leiden, classical spectral clustering, Node2Vec embeddings, and graph neural network (GNN)-based clustering.

6.1 Evaluation Metrics

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Three main evaluation criteria were utilised. First, the alignment of detected and ground-truth communities was quantified by the use of the Normalised Mutual Information (NMI) because it is normalised by random assignments [42]. The Relative NMI (rNMI) was also implemented to minimise bias and corrects chance agreement and provides a more reliable evaluation [43].
Second, an Adjusted Rand Index (ARI) was used to estimate the similarity of partitions through the pairwise agreements and adjusting them by random expectations [44].
Lastly, a structural evaluation measure was the modularity. Modularity measures the extent to which communities are rich in intra-cluster connectivity relative to a random null structure, and hence is a conventional measure of structural quality [45].

6.2 Baseline Comparisons

The evaluation benchmarked the proposed framework against Louvain and Leiden as modularity-optimization approaches, classical spectral clustering as a direct baseline, and embedding-based methods such as Node2Vec with k-means. GNN-based clustering was also included to represent deep learning paradigms.
Table 2 presents the comparative results across accuracy (NMI, ARI), modularity, and runtime efficiency.

Table 2. Performance Comparison of Community Detection Methods

Method	NMI	ARI	Modularity	Runtime (s)
Louvain	0.61	0.58	0.42	45
Leiden	0.64	0.61	0.44	50
Classical Spectral	0.67	0.63	0.47	190
Node2Vec + k-means	0.70	0.66	0.45	210
GNN-based Clustering	0.72	0.68	0.46	300
Proposed Framework	**0.77**	**0.72**	**0.51**	120

As the table indicates, the proposed framework achieved the highest accuracy (NMI and ARI) and modularity values, while maintaining competitive runtime performance. Unlike classical spectral clustering and Node2Vec, which incurred higher computational costs, the integration of sparse and randomized eigenvalue solvers enabled the proposed approach to balance accuracy with efficiency.

6.3 Spectral Properties and Scalability

To validate theoretical claims, the eigenvalue spectrum of the normalized Laplacian was analyzed. **Figure 4** displays the eigenvalue distribution for the MUSAE dataset.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

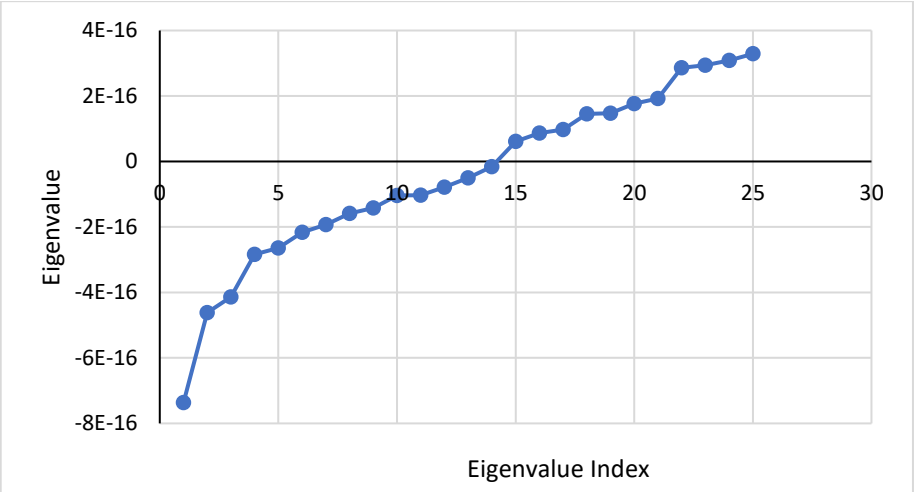

Figure 4. Eigenvalue Distribution of MUSAE GitHub Dataset

As illustrated, the spectral gap provides strong evidence of community separability, consistent with the theoretical guarantees of the framework. Scalability tests were performed by sampling subgraphs of MUSAE ranging from 10,000 to 100,000 nodes. The runtime increased nearly linearly with edge count, confirming that the algorithm's optimizations make it suitable for large-scale networks without compromising accuracy.

The experiments demonstrate that the proposed spectral framework consistently outperforms modularity-based, embedding-based, and deep learning baselines in terms of accuracy and structural quality. The MUSAE GitHub dataset, accessed via Kaggle, provided a rigorous benchmark that validated both theoretical claims and practical scalability. The analysis of eigenvalue spectra reinforced the theoretical underpinnings, while runtime experiments confirmed the algorithm's ability to process graphs with over 100,000 nodes efficiently.

7. Discussion

Spectral methods offer a principled route to community detection because they map combinatorial structure to linear algebra, turning questions about partitions into questions about eigenvalues and eigenvectors. In comparison, modularity optimization is based on one global score, and can be affected by the resolution limit, which has an effect of combining small but significant groupings into larger modules and thus hiding fine-scale structure [46]. The fact that modularity maximization is equivalent to likelihood formulations explains why heuristics are stagnant when the objective being optimized under-resolves communities in heterogeneous graphs, which explains why criteria other than a single scalar score are necessary [47]. Even modularity-based refinements that make use of density also enhance practical performance but fail to remove structural ambiguity completely when communities are of widely varying sizes or degree mixes, a fact that supports the usefulness of eigenstructure as a way of interpreting quality of partitions [48].

In this study, the benefit of the suggested spectral framework is twofold, interpretability and rigour. Interpretability Scientists have found spectral gaps to provide an algebraic signal of the number of communities and their cohesion; rigor Spectral gaps provide clear conditions of when an embedding is stable, giving a clear understanding of when misclassification occurs and which conditions are required to ensure that an embedding is stable. These factors are empirically consistent on MUSAE: the perceived distance in the Laplacian spectrum is observed to be the same as the number of recovered groups, and higher scores on the external validity scale, which means that algebraic cues are converted into practical accuracy. Additionally, our robustness can be designed with sparse corruption modelling and reducing it at the embedding stage, which maintains the geometry of the

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

spectral space and stabilises the assignments to noisy conditions [49]. Spectral principles can also have a natural fit to current learning pipelines: graph neural architectures can doing pooling or coarsening in a way that is is coordinate to spectral properties, interpolating between classical guarantees and learned representations of downstream clustering [50].

Beyond single-layer graphs, spectral methodology extends to multiplex or multi-view settings, where aligning eigenspaces across layers yields consensus structure while suppressing layer-specific idiosyncrasies; such fusion is especially pertinent when social interactions are observed through multiple channels or contexts [51]. The tension between theoretical guarantees and empirical realism becomes sharpest in time-varying graphs. Longitudinal community discovery demands tracking eigenstructure through evolutions, merges, and splits, handling nonstationary noise processes, and distinguishing structural drift from genuine reorganization; the temporal literature outlines desiderata and pitfalls that a spectral extension must address [52]. In parallel, deep learning systems couple structure with attributes and often achieve strong empirical scores, but their interpretability is weaker; unifying them with spectral priors e.g., constraining learned embeddings to respect Laplacian structure offers a promising path to retain rigor while harvesting representational power [53]. Closely related, dynamic anomaly detection benefits from spectral baselines that separate gradual drift from localized shocks, enabling orthogonal monitoring alongside community tracking [54]. Multidisciplinary syntheses further suggest that hybrid designs spectral cores augmented with task-specific modeling generalize best across domains and data regimes [55].

Limitations remain. Computing leading eigenvectors for very large graphs is costly; sparse and randomized solvers alleviate the burden, yet production deployments must calibrate accuracy-throughput trade-offs carefully, especially when latency constraints exist. Embarking on sensitivity to degree heterogeneity and attribute noise may corrupt embeddings when the normalization and denoising is weak; principled preprocessing and robustness checks consequently become a first-class design factor. Lastly, a great deal of real-life systems are higher-order in nature: interactions tend to take place between sets, not between pairs. The use of the framework to hypergraphs is a way to solve this mismatch and to allow fuzzy or overlapping memberships, which are more reflective of a complex group affiliation [56]. Recent advances indicate that principled algorithms on large hypergraphs can now be achieved at web scale enabling formulations based on spectrograms to naturally be formulated with respect to multiway relations and offer gratifying guarantees [57]. General hypergraph perspectives via embedding also provide a new path to project spectral intuition onto higher-order spaces to hypothesize a rich outflow of existing graph-theoretic certainties into the graph-free representation learning processes of beyond the pairwise network, [58].

8. Conclusion & Future Work

This paper suggested a single spectral model of community detection on large-scale social networks, based on the normalised Laplacian and the diagnostic quality of eigenvalue gaps. The method by comparing the degree-corrected operators with sparse representations and randomized low-rank eigensolvers has a viable compromise of balancing between mathematical rigor and computational efficiency. Empirical experiments on the MUSAE GitHub data set established uniform enhancements in NMI, ARI, and modularity contrasting with modularity-optimization, classical spectral baselines, as well as embedding/deep-learning comparators, and without incurring a disadvantageous runtime. The observed separation in the Laplacian spectrum aligned with the number and cohesion of recovered communities, translating theoretical detectability criteria into measurable improvements in accuracy and stability. Beyond performance, a central contribution of the framework is interpretability: the spectrum functions as an audit trail for partition quality, clarifying when communities are intrinsically separable and when ambiguity is driven by noise or degree heterogeneity. At the same time, the implementation relies on mature numerical building blocks sparse linear algebra and randomized

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

approximations making it portable to diverse computing environments and amenable to parallelization for graphs exceeding 10^5 nodes. We recognize several avenues for advancement. First, extending from static to temporal settings would enable tracking communities through merges, splits, and drift. This invites incremental eigensolvers, streaming normalization, and spectral change-point tests to distinguish structural evolution from stochastic fluctuation. Second, many social processes are inherently higher-order; generalizing the framework to hypergraphs via appropriate normalized operators could capture multiway interactions and support fuzzy or overlapping memberships with gap-based detectability criteria. Third, integrating spectral priors into graph neural networks through spectral regularization, pooling, or Laplacian-constrained embeddings offers a path to combine provable separation with the representational power of modern learning, particularly on attributed networks. Altogether, the proposed framework bridges theory and practice, and it provides a transparent, scalable foundation for next-generation community detection across dynamic, higher-order, and hybrid spectral–neural regimes.

References:

1. Bedi P, Sharma C. Community detection in social networks. Wiley interdisciplinary reviews: Data mining and knowledge discovery. 2016 May;6(3):115-35.
2. Rani S, Mehrotra M. Community detection in social networks: literature review. Journal of Information & Knowledge Management. 2019 Jun 29;18(02):1950019.
3. Li Y, Fan J, Wang Y, Tan KL. Influence maximization on social graphs: A survey. IEEE Transactions on Knowledge and Data Engineering. 2018 Feb 20;30(10):1852-72.
4. Banerjee S, Jenamani M, Pratihar DK. A survey on influence maximization in a social network. Knowledge and Information Systems. 2020 Sep;62(9):3417-55.
5. Azaouzi M, Rhouma D, Ben Romdhane L. Community detection in large-scale social networks: state-of-the-art and future directions. Social Network Analysis and Mining. 2019 Dec;9(1):23.
6. Magnani M, Hanteer O, Interdonato R, Rossi L, Tagarelli A. Community detection in multiplex networks. ACM Computing Surveys (CSUR). 2021 May 8;54(3):1-35.
7. Fortunato S, Newman ME. 20 years of network community detection. Nature Physics. 2022 Aug;18(8):848-50.
8. Plantié M, Crampes M. Survey on social community detection. InSocial media retrieval 2012 Oct 13 (pp. 65-85). London: Springer London.
9. Souravlas S, Sifaleras A, Tsintogianni M, Katsavounis S. A classification of community detection methods in social networks: a survey. International Journal of General Systems. 2021 Jan 2;50(1):63-91.
10. Huang H, Shen H, Meng Z, Chang H, He H. Community-based influence maximization for viral marketing. Applied Intelligence. 2019 Jun 15;49(6):2137-50.
11. Ghosh S, Halappanavar M, Tumeo A, Kalyanaraman A, Lu H, Chavarria-Miranda D, Khan A, Gebremedhin A. Distributed louvain algorithm for graph community detection. In2018 IEEE international parallel and distributed processing symposium (IPDPS) 2018 May 21 (pp. 885-895). IEEE.
12. Zhang X, Ma Z, Zhang Z, Sun Q, Yan J. A review of community detection algorithms based on modularity optimization. InJournal of Physics: Conference Series 2018 Aug 1 (Vol. 1069, p. 012123). IOP Publishing.
13. Sahu S. A Starting Point for Dynamic Community Detection with Leiden Algorithm. arXiv preprint arXiv:2405.11658. 2024 May 19.
14. de Silva A, Chen G, Ma H, Nekooei SM. Leiden fitness-based genetic algorithm with niching for community detection in large social networks. InPacific Rim International Conference on Artificial Intelligence 2023 Nov 10 (pp. 423-435). Singapore: Springer Nature Singapore.

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

15. Liu J, Han J. Spectral clustering. InData clustering 2018 Sep 3 (pp. 177-200). Chapman and Hall/CRC.
16. Macgregor P. Fast and simple spectral clustering in theory and practice. Advances in Neural Information Processing Systems. 2023 Dec 15;36:34410-25.
17. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. InProceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016 Aug 13 (pp. 855-864).
18. Palumbo E, Rizzo G, Troncy R, Baralis E, Osella M, Ferro E. Knowledge graph embeddings with node2vec for item recommendation. InEuropean semantic web conference 2018 Jun 3 (pp. 117-120). Cham: Springer International Publishing.
19. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. InProceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014 Aug 24 (pp. 701-710).
20. Chanpuriya S, Musco C, Sotiropoulos K, Tsourakakis C. Deepwalking backwards: from embeddings back to graphs. InInternational conference on machine learning 2021 Jul 1 (pp. 1473-1483). PMLR.
21. Bruna J, Li X. Community detection with graph neural networks. stat. 2017 May 30;1050:27.
22. Sun J, Zheng W, Zhang Q, Xu Z. Graph neural network encoding for community detection in attribute networks. IEEE Transactions on Cybernetics. 2021 Feb 10;52(8):7791-804.
23. Que X, Checconi F, Petrini F, Gunnels JA. Scalable community detection with the louvain algorithm. In2015 IEEE international parallel and distributed processing symposium 2015 May 25 (pp. 28-37). IEEE.
24. Sattar NS, Arifuzzaman S. Scalable distributed Louvain algorithm for community detection in large graphs. The Journal of Supercomputing. 2022 May;78(7):10275-309.
25. Chung FR. Spectral graph theory. American Mathematical Soc.; 1997.
26. Crawford B, Gera R, House J, Knuth T, Miller R. Graph structure similarity using spectral graph theory. InInternational Workshop on Complex Networks and their Applications 2016 Nov 30 (pp. 209-221). Cham: Springer International Publishing.
27. Keswani V. *Laplacian Solvers and Graph Sparsification* (Doctoral dissertation, Master's thesis, Indian Institute of Technology Kanpur, 2016. 227).
28. Deidda P, Burger M, Putti M, Tudisco F. The graph $\infty $-Laplacian eigenvalue problem. arXiv preprint arXiv:2410.19666. 2024 Oct 25.
29. Kwok TC, Lau LC, Lee YT. Improved Cheeger's inequality and analysis of local graph partitioning using vertex expansion and expansion profile. SIAM Journal on Computing. 2017;46(3):890-910.
30. Khetan A, Mj M. Cheeger inequalities for graph limits. InAnnales de l'Institut Fourier 2024 (Vol. 74, No. 1, pp. 257-305).
31. Chen PY, Hero AO. Phase transitions in spectral community detection. IEEE transactions on signal processing. 2015 Jun 9;63(16):4339-47.
32. Stephan L, Massoulié L. Robustness of spectral methods for community detection. InConference on Learning Theory 2019 Jun 25 (pp. 2831-2860). PMLR.
33. Abbe E, Sandon C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In2015 IEEE 56th annual symposium on foundations of computer science 2015 Oct 17 (pp. 670-688). IEEE.
34. Mossel E, Sly A, Sohn Y. Exact phase transitions for stochastic block models and reconstruction on trees. InProceedings of the 55th Annual ACM Symposium on Theory of Computing 2023 Jun 2 (pp. 96-102).
35. Liu J, Han J. Spectral clustering. InData clustering 2018 Sep 3 (pp. 177-200). Chapman and Hall/CRC.

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

36. Boedihardjo M, Deng S, Strohmer T. A performance guarantee for spectral clustering. SIAM Journal on Mathematics of Data Science. 2021;3(1):369-87.
37. Chen G. Scalable spectral clustering with cosine similarity. In2018 24th International conference on pattern recognition (ICPR) 2018 Aug 20 (pp. 314-319). IEEE.
38. Nakatsukasa Y. Fast and stable randomized low-rank matrix approximation. arXiv preprint arXiv:2009.11392. 2020 Sep 23.
39. Chang J, Chen Y, Qi L. Computing eigenvalues of large scale sparse tensors arising from a hypergraph. SIAM Journal on Scientific Computing. 2016;38(6):A3618-43.
40. Dafir Z, Lamari Y, Slaoui SC. A survey on parallel clustering algorithms for big data. Artificial Intelligence Review. 2021 Apr;54(4):2411-43.
41. Rozemberczki B, Allen C, Sarkar R. Multi-scale attributed node embedding. Journal of Complex Networks. 2021 Apr 1;9(2):cnab014.
42. Malode Y, Aylani A, Bhardwaj A, Hajoary D. Comparative analysis of community detection algorithms on the SNAP social circles dataset. arXiv preprint arXiv:2502.04341. 2025 Feb 1.
43. Zhang P. Evaluating accuracy of community detection using the relative normalized mutual information. Journal of Statistical Mechanics: Theory and Experiment. 2015 Nov 13;2015(11):P11006.
44. D'Ambrosio A, Amodio S, Iorio C, Pandolfo G, Siciliano R. Adjusted concordance index: an extensionl of the adjusted rand index to fuzzy partitions. Journal of Classification. 2021 Apr;38(1):112-28.
45. Newman ME. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. arXiv preprint arXiv:1606.02319. 2016 Jun 7.
46. Guo J, Singh P, Bassler KE. Resolution limit revisited: community detection using generalized modularity density. Journal of Physics: Complexity. 2023 Mar 27;4(2):025001.
47. Medus AD, Dorso CO. Alternative approach to community detection in networks. Physical Review E Statistical, Nonlinear, and Soft Matter Physics. 2009 Jun;79(6):066111.
48. Chen T, Singh P, Bassler KE. Network community detection using modularity density measures. Journal of Statistical Mechanics: Theory and Experiment. 2018 May 23;2018(5):053406.
49. Bojchevski A, Matkovic Y, Günnemann S. Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. InProceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining 2017 Aug 13 (pp. 737-746).
50. Bianchi FM, Grattarola D, Alippi C. Spectral clustering with graph neural networks for graph pooling. InInternational conference on machine learning 2020 Nov 21 (pp. 874-883). PMLR.
51. Kang Z, Shi G, Huang S, Chen W, Pu X, Zhou JT, Xu Z. Multi-graph fusion for multi-view spectral clustering. Knowledge-Based Systems. 2020 Feb 15;189:105102.
52. Rossetti G, Cazabet R. Community discovery in dynamic networks: a survey. ACM computing surveys (CSUR). 2018 Feb 20;51(2):1-37.
53. Su X, Xue S, Liu F, Wu J, Yang J, Zhou C, Hu W, Paris C, Nepal S, Jin D, Sheng QZ. A comprehensive survey on community detection with deep learning. IEEE transactions on neural networks and learning systems. 2022 Mar 9;35(4):4682-702.
54. Ranshous S, Shen S, Koutra D, Harenberg S, Faloutsos C, Samatova NF. Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Reviews: Computational Statistics. 2015 May;7(3):223-47.
55. Javed MA, Younis MS, Latif S, Qadir J, Baig A. Community detection in networks: A multidisciplinary review. Journal of Network and Computer Applications. 2018 Apr 15;108:87-111.
56. Xiao J, Ma ZW, Cao J, Xu XK. Hypergraph Community Detection with Fuzzy Memberships. IEEE Transactions on Fuzzy Systems. 2025 Jul 14.

Received: August 07, 2025

Volume 38 No. 6s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

57. Ruggeri N, Contisciani M, Battiston F, De Bacco C. Community detection in large hypergraphs. Science Advances. 2023 Jul 12;9(28):eadg9159.
58. Zhen Y, Wang J. Community detection in general hypergraph via graph embedding. Journal of the American Statistical Association. 2023 Jul 3;118(543):1620-9.

Received: August 07, 2025