IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/44AI/2024/4455724/
Vol. 2024 (22), Article ID 4455724

A Hybrid Approach for Community and Anomaly Detection in Social Networks*

Sarah ALHARBI and Hedia ZARDI

Department of Computer Science, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia

Correspondence should be addressed to: Sarah ALHARBI, 411207154@qu.edu.sa

* Presented at the 44th IBIMA International Conference, 27-28 November 2024 Granada, Spain

Abstract

Detecting communities and anomalies in social networks is critical for understanding complex network structures and identifying irregular behaviors. Existing methods often address these tasks separately, overlooking their interdependence. This study addresses this gap by proposing a hybrid approach that integrates modularity-based community detection with structural and attribute-based similarity measures for anomaly detection. The framework leverages modularity optimization to identify cohesive communities and employs cosine similarity to evaluate attribute alignment for detecting anomalies. Evaluations on synthetic and real-world datasets, including the Cora and Cite seer citation networks, demonstrate the framework's superior performance over state-of-the-art techniques, achieving higher precision, recall, and F1 scores. These findings underscore the method's robustness, scalability, and potential for applications in social network analysis, fraud detection, and security.

Keywords: Community Detection, Anomaly Detection, Social Network Analysis (SNA), Modularity Optimization, Structural Similarity, Attribute-Based Similarity, Graph Theory, Attributed Graphs

Introduction

Online social networks (OSNs) have become an integral part of modern communication, collaboration, and information exchange. These networks facilitate interactions among millions of users daily, creating complex structures that reflect intricate social dynamics (Fortunato and Hric 2016), (Rossetti and Cazabet 2018). Understanding these structures requires advanced analytical approaches to uncover the underlying patterns and relationships within these networks (Zhang et al. 2022), (Ackoff 1961). Two critical tasks in this field are community detection and anomaly detection. Community detection focuses on identifying groups of nodes that are tightly connected and share common interests or behaviors (Chen et al. 2021), (Wu et al. 2021). Anomaly detection, on the other hand, seeks to identify nodes or edges that deviate significantly from normal patterns, which may indicate fraudulent activity or structural inconsistencies (Gao et al. 2020), (Akoglu and Tong 2023).

As OSNs continue to grow in scale and complexity, addressing these tasks simultaneously has become essential. Communities provide valuable context for identifying anomalies, while anomalies can alter or reveal hidden community structures (Akoglu and Tong 2023). This paper proposes a hybrid approach that integrates graph structure and node attributes to simultaneously detect communities and anomalies. By combining modularity-based community detection with structural and attribute-based similarity measures, the proposed method achieves high accuracy and robustness. Evaluated on both synthetic and real-world datasets, this approach demonstrates its ability to effectively analyze complex networks and detect anomalies, offering a reliable solution for advancing social network analysis.

Cite this Article as: Sarah ALHARBI and Hedia ZARDI, Vol. 2024 (22) "A Hybrid Approach for Community and Anomaly Detection in Social Networks " Communications of International Proceedings, Vol. 2024 (22), Article ID 4455724, https://doi.org/10.5171/2024.4455724

Related Work

Community detection and anomaly detection are two critical tasks in social network analysis that have received significant attention in recent years. Community detection focuses on identifying groups of nodes that are more densely connected internally than with the rest of the network. Traditional approaches, such as modularity optimization, evaluate the quality of community partitions by comparing the density of intra-community edges against random graphs (Fortunato and Hric 2016). Spectral clustering is another widely adopted method that leverages the eigenvalues of graph Laplacians to identify clusters (Rossetti and Cazabet 2018). Recently, graph neural networks (GNNs) have emerged as powerful tools for community detection, as they integrate both structural and attribute-based information, allowing for more precise identification of community structures (Zhang et al. 2022). An example of this innovation is edge-enhanced graph neural networks, which explicitly incorporate edge features, improving the detection of cohesive groups (Chen et al. 2021).

Anomaly detection, on the other hand, aims to identify nodes or edges that deviate from expected patterns. Structural anomaly detection focuses on the topology of the graph, identifying irregularities such as nodes that act as bridges between distinct communities or those with unexpected degrees (Ackoff 1961). Attributed anomaly detection extends this by utilizing node and edge features to identify anomalies that may not be evident from the graph structure alone (Gao et al. 2020). Recent advances include methods such as multiple graph attention networks, which process both structural and attribute-based data to detect anomalies with higher precision (Liu et al. 2022).

Hybrid approaches that combine community detection and anomaly detection have recently gained traction, addressing the interdependence between these tasks. Communities often provide essential context for identifying anomalies, while anomalies can reveal hidden or disrupted community structures (Akoglu and Tong 2023). Techniques such as graph embedding, including node2vec, and graph convolutional networks (GCNs) have shown promise in learning latent node representations to jointly detect communities and anomalies (Wu et al. 2021). Despite these advancements, hybrid methods still face challenges, particularly with scalability and noise sensitivity in large-scale networks.

This thesis builds upon these recent advancements by proposing a hybrid approach that integrates modularity-based community detection with structural and attribute-based similarity measures for anomaly detection. By leveraging both the topological and attribute information of networks, the proposed framework aims to address the limitations of previous methods and provide an efficient, unified solution for analyzing complex networks.

Methodology

The proposed approach integrates community detection and anomaly detection into a unified framework, leveraging both structural properties and attribute information of networks. This methodology is organized into four sequential steps: seed selection, similarity measures, anomaly score calculation, and node classification. Each step ensures that the network is comprehensively analyzed for cohesive communities and anomalous behaviors.

Seed Node Selection

Seed selection is the initial step, identifying influential nodes that serve as starting points for community formation. Nodes are ranked based on a combination of centrality measures:

- **Degree Centrality:** Evaluates the importance of a node based on the number of direct connections (Freeman 1978).
- **Closeness Centrality:** Measures how accessible a node is to all other nodes in the network (Newman 2018).
- **Betweenness Centrality:** Quantifies how often a node acts as a bridge along the shortest path between other nodes (Brandes 2001).
- **Eigenvector Centrality:** Identifies nodes connected to other highly influential nodes (Bonacich 1987).
- **PageRank:** Evaluates node importance based on a recursive scoring of its neighbors (Brin and Page 1998).

Nodes with the highest combined centrality scores are selected as seeds, ensuring that communities grow around well connected and influential nodes.

A. Structural and Attribute Similarity

To evaluate a node's compatibility with a community, two types of similarity measures are used:

1) Structural Similarity:

Modularity is used to measure the quality of community partitions by comparing observed connections within communities to those expected in a random graph (Newman 2006). When a node *v* is added to a community *C*, the modularity gain ΔQ is computed as:

$$\Delta Q = \frac{\text{Internal Edges} \; - \; \text{Expected Edges}}{\text{Total Possible Edges}}$$

The modularity gain (ΔQ) evaluates the improvement in modularity when a node is added to a community. It measures how well the node fits into the community based on the density of intra-community connections. The value of ΔQ is interpreted as follows:

- $\Delta Q > 0$: Indicates that adding the node improves the modularity of the community, suggesting a strong fit.
- $\Delta Q = 0$: Indicates no change in modularity, implying that the node neither benefits nor harms the community's structure.
- $\Delta Q < 0$: Indicates a reduction in modularity, suggesting that the node does not align well with the community.

For normalized modularity gain calculations, ΔQ typically ranges between 0 and 1, with higher values indicating better integration of the node into the community. This property ensures consistency with other measures, such as Cosine Similarity, used in the framework.

Attribute-Based Similarity:

The Cosine Similarity is used to measure the similarity of node attributes (Singhal 2001):

$$\text{CosSim}(u, v) \; = \frac{\mathbf{A}_u \cdot \mathbf{A}_v}{\|\mathbf{A}_u\| \|\mathbf{A}_v\|},$$

where $\mathbf{A}_u$ and $\mathbf{A}_v$ are the attribute vectors of nodes *u* and *v*. Nodes with low cosine similarity are flagged as potentially anomalous.

The Cosine Similarity formula for a node *v* to a community *C* is defined as:

$$\text{CosSim}(v, C) \; = \frac{\mathbf{A}_v \; \cdot \; \mathbf{A}_C}{\|\mathbf{A}_v\| \|\mathbf{A}_C\|},$$

where:

- $\mathbf{A}_v$ is the **attribute vector** of the node *v*,
- $\mathbf{A}_C$ is the **aggregated attribute vector** of the community *C*, representing the combined attributes of all nodes in *C*,
- $\mathbf{A}_v \; \cdot \; \mathbf{A}_C$ is the **dot product** of the node's and community's attribute vectors,
- $\|\mathbf{A}_v\|$ is the **Euclidean norm** (magnitude) of the node's attribute vector, calculated as:

$$\|\mathbf{A}_v\| \; = \sqrt{\sum_i A_{v,i}^2},$$

- $\|\mathbf{A}_C\|$ is the **Euclidean norm** (magnitude) of the aggregated community attribute vector, calculated similarly.

The Cosine Similarity measures the alignment of a node's attributes with the overall attributes of the community. A value close to 1 indicates strong similarity, while a value close to 0 indicates weak similarity. In other words, a note explicitly states that $CosSim(v, C)$ is bounded between 0 (completely dissimilar) and 1 (completely similar).

Anomaly Score Calculation

In this step, two scores are calculated for each node to determine its compatibility with a community or its likelihood of being an anomaly:

- **Adding Score** quantifies how well a node v fits into a community C by combining two important factors:
  – ΔQ: The modularity gain, which measures the improvement in the structural quality of the community if the node v is added. A higher ΔQ indicates that the node strongly aligns with the structural properties of the community.
  – $\mathrm{CosSim}(v, C)$: The cosine similarity, which evaluates how closely the attributes of the node v match the aggregate attributes of the community C. A higher $\mathrm{CosSim}(v, C)$ reflects better compatibility between the node's attributes and the community's overall characteristics.

The Adding Score is calculated as:

Adding Score = $\alpha \cdot \Delta Q + (1 - \alpha) \cdot \mathrm{CosSim}(v, C)$, where:

– $\alpha \in [0, 1]$ is a weighting parameter that balances the contributions of ΔQ and $\mathrm{CosSim}(v, C)$.

– A higher Adding Score indicates that the node v is a strong candidate for inclusion in the community C, based on both structural and attribute compatibility.

This combined approach ensures that nodes are evaluated holistically, taking into account both their structural role within the graph and their attribute alignment with the community. The flexibility provided by the parameter α allows the score to be tailored to specific networks, emphasizing structural properties (ΔQ) or attribute similarity ($\mathrm{CosSim}(v, C)$) as needed.

- **Rejection Score** evaluates how poorly a node v aligns with a community C. It is calculated as:

Rejection Score = $\alpha \cdot (-\Delta Q) + (1-\alpha) \cdot (1 - \mathrm{CosSim}(v, C))$, where:

– $(-\Delta Q)$: The negative modularity gain, which penalizes nodes that reduce the structural quality of the community if added. A larger negative ΔQ contributes to a higher rejection score, indicating poor structural compatibility.

– $(1 - \mathrm{CosSim}(v, C))$: The dissimilarity of the node's attributes to the aggregate attributes of the community. A higher value reflects greater deviation from the community's characteristics.

– $\alpha \in [0, 1]$: A weighting parameter that balances the contributions of structural $(-\Delta Q)$ and attribute based $(1 - \mathrm{CosSim}(v, C))$ dissimilarities.

The Rejection Score prioritizes nodes that negatively impact the community's modularity or significantly deviate from its attributes. A higher score suggests that the node v is more likely to be an anomaly, as it neither fits structurally nor aligns in terms of attributes with the community.

B. Node Classification

The node classification process determines whether a node v should be added to a community C or flagged as an anomaly. This decision is based on two scores: Adding Score and Rejection Score.

The decision rule is made as follow:

- If the **Adding Score** is higher, the node is added to the community.
- If the **Rejection Score** is higher, the node is flagged as an anomaly.

In conclusion, this methodology integrates community detection and anomaly detection into a unified framework. By leveraging centrality measures, modularity optimization, and attribute similarity, the proposed approach ensures robust and scalable analysis of large and complex networks.

Experimental Evaluation

The performance of the proposed method is compared with two baseline models, which are introduced in this section. Additionally, we describe the evaluation metrics used for the assessment and provide a detailed description of the datasets.

A. Comparative Models for Community and Anomaly Detection

1) *Bayesian Robust Attributed Graph Clustering:* BRAC is a probabilistic framework that integrates graph structure and node attributes for clustering attributed graphs. It employs Bayesian modeling to robustly assign nodes to communities while handling noisy and incomplete data. This method balances the influence of structural and attribute information, making it effective in uncovering latent community structures in noisy environments (Zhang et al. 2021).

2) *Anomaly and Community Detection:* ACD is an integrated approach designed to detect both anomalies and communities in attributed graphs. It jointly optimizes community detection and anomaly identification by leveraging node connectivity and attribute consistency. ACD excels in identifying anomalies that disrupt community structures while maintaining a high precision in community detection, even in complex, large-scale networks (Gao et al. 2022).

B. Evaluation Metrics

The performance of the proposed method is evaluated using the following metrics:

a) Normalized Mutual Information (NMI): Normalized Mutual Information measures the similarity between the detected community structure and the ground-truth community labels. It is defined as:

$$NMI(U,V) = \frac{2 \cdot I(U,V)}{H(U) + H(V)},$$

where:

- U and V are the sets of ground-truth and detected communities, respectively,
- $I(U,V)$ is the mutual information between U and V,
- $H(U)$ and $H(V)$ are the entropies of U and V.

NMI ranges between 0 and 1, where 1 indicates perfect alignment between the detected and true community structures.

b) Precision: Precision measures the proportion of correctly identified anomalies among all detected anomalies. It is defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}.$$

A higher precision indicates fewer false positives.

c) Recall: Recall evaluates the proportion of actual anomalies that are correctly identified. It is given by:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}.$$

A higher recall indicates fewer missed anomalies.

d) F1 Score: The F1 Score is the harmonic mean of precision and recall, providing a balanced measure of the two. It is defined as:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

The F1 Score ranges between 0 and 1, where 1 indicates perfect precision and recall.

Table 1: EVALUATION METRICS

Metric	Definition
NMI	Agreement between detected communities and ground truth.
Precision	Proportion of correctly identified anomalies.
Recall	Proportion of actual anomalies detected.
F1 Score	Harmonic mean of Precision and Recall.

C. Description of Datasets

The evaluation of the proposed method is conducted using three datasets: one synthetic and two real-world datasets. These datasets provide a range of structural and attribute-based challenges to test the method's robustness.

a. Synthetic Dataset: The synthetic dataset is generated using the Stochastic Block Model (SBM), a widely used model for creating networks with predefined community structures. It consists of:

- **Nodes:** 500 nodes divided into 3 distinct communities.
- **Edges:** Edges are generated based on intra-community and inter-community probabilities, ensuring clear community boundaries.
- **Anomalies:** Controlled anomalies are introduced by adding or removing edges and modifying node attributes to simulate outliers.

This dataset evaluates the method's ability to detect well defined communities and identify injected anomalies.

b. Cora Citation Network: The Cora dataset is a real-world attributed graph containing scientific publications. Each node represents a publication, and edges represent citation relationships. The dataset is characterized by:

- **Nodes:** 2,708 publications.
- **Edges:** 5,429 citation links.
- **Attributes:** Each node has a 1,433-dimensional binary feature vector indicating the presence of specific keywords in the publication.
- **Classes:** Publications are classified into 7 research topics, providing ground-truth labels for community detection.

This dataset tests the method's ability to leverage both structural and attribute information for community detection and anomaly identification.

c. Citeseer Citation Network: The Citeseer dataset is another real-world attributed graph similar to Cora but with greater structural complexity. It contains:

- **Nodes:** 3,312 publications.
- **Edges:** 4,732 citation links.
- **Attributes:** Each node has a 3,703-dimensional binary feature vector based on keywords.
- **Classes:** Publications are grouped into 6 research topics.

This dataset is used to evaluate the scalability and effectiveness of the proposed method on larger, more challenging networks.

Table 2: DATASET STATISTICS

Dataset	Nodes	Edges
Synthetic Dataset	500	1,200
Cora Citation Network	2,708	5,429
Citeseer Citation Network	3,312	4,732

The combination of synthetic and real-world datasets ensures a comprehensive evaluation of the proposed method. The synthetic dataset provides controlled testing conditions, while the Cora and Citeseer networks simulate real-world challenges in community and anomaly detection.

Results And Discussion

In this section, we analyze the performance of the proposed method on each dataset individually, comparing it with baseline models.

A. Synthetic Dataset

The Synthetic Dataset was generated using the Stochastic Block Model (SBM), which provides a controlled environment for evaluating the robustness of community detection and anomaly identification.

As shown in Figure 1, the proposed method consistently outperformed BRAC and ACD across all evaluation metrics:

- **NMI:** The proposed method achieved an NMI of 0.85, surpassing BRAC (0.78) and ACD (0.80), demonstrating superior community detection alignment with ground truth.
- **F1 Score:** The F1 score of 0.89 highlights the method's ability to balance precision and recall effectively, outperforming BRAC (0.85) and ACD (0.85).

The controlled nature of this dataset validates the proposed method's integration of modularity gain and attribute similarity as a robust approach to detecting both communities and anomalies.

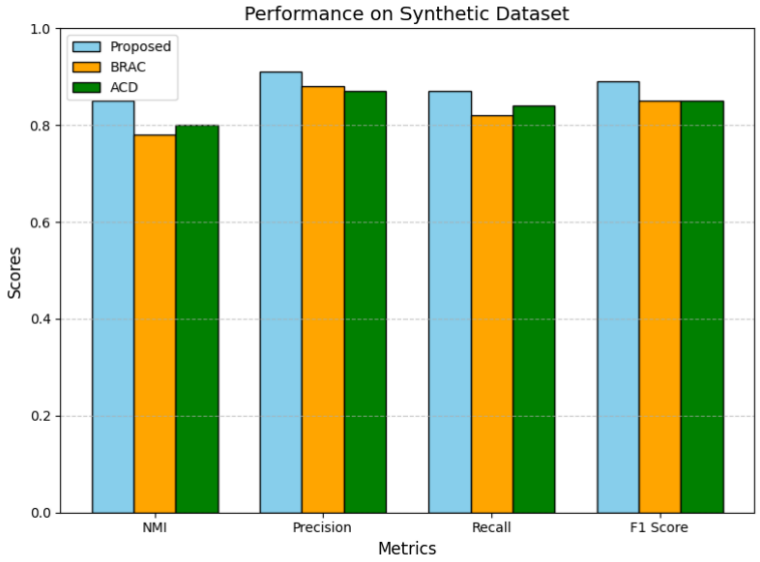

Fig 1. Performance comparison on the Synthetic Dataset.

Cora Citation Network

The Cora dataset represents a real-world citation network with rich attribute information. The performance of the proposed method on this dataset is summarized in Figure 2.

- **NMI:** The proposed method achieved an NMI of 0.72, outperforming BRAC (0.70) and ACD (0.71), reflecting strong alignment with ground-truth communities.
- **F1 Score:** With an F1 score of 0.85, the proposed method demonstrated a balance between high precision and recall, compared to BRAC (0.82) and ACD (0.84).

This dataset highlights the method's ability to integrate structural and attribute information effectively in a real-world scenario.

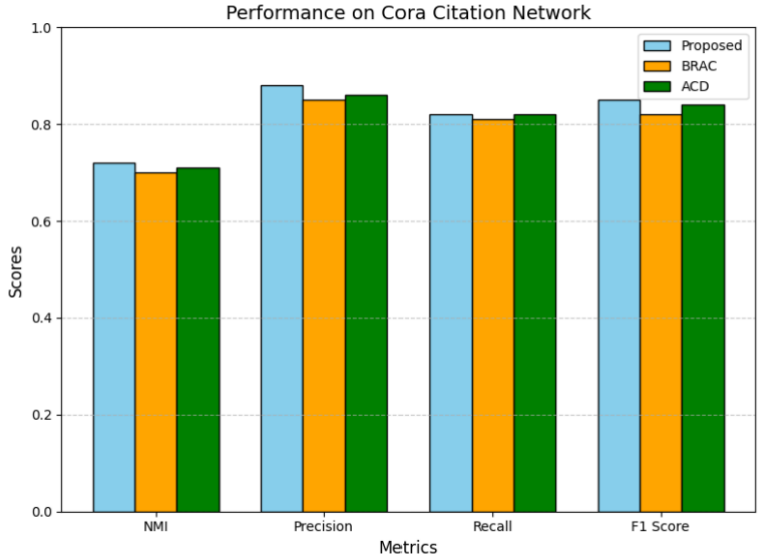

Fig 2. Performance comparison on the Cora Citation Network.

B. Citeseer Citation Network

The Citeseer dataset, with its more complex structure and larger size, tests the scalability of the proposed method. The results are shown in Figure 3.

- **NMI:** The proposed method achieved an NMI of 0.68, slightly higher than BRAC (0.65) and ACD (0.66), showcasing its scalability to larger networks.
- **F1 Score:** The F1 score of 0.85 indicates consistent performance, outperforming BRAC (0.82) and ACD (0.83).

These results confirm the proposed method's robustness in handling larger and more complex datasets while maintaining high performance across evaluation metrics.

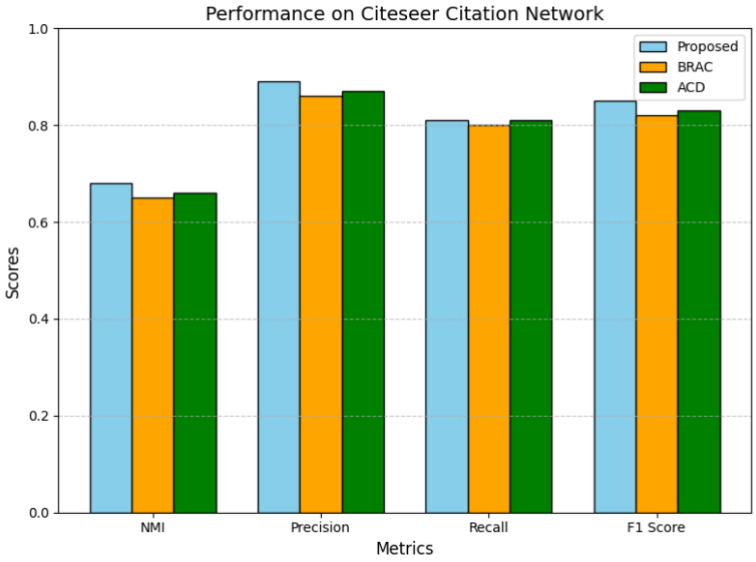

Fig 3. Performance comparison on the Citeseer Citation Network.

Conclusion

This paper introduced a hybrid framework for community and anomaly detection in social networks, integrating modularity gain (ΔQ) with attribute similarity (CosSim) to evaluate nodes holistically. The framework was evaluated on synthetic and real-world datasets, demonstrating superior performance compared to state-of-the-art models (BRAC and ACD).

Key findings include:

- The proposed method achieved the highest F1 scores on all datasets, highlighting its robustness in balancing precision and recall.
- Consistent performance on the Cora and Citeseer datasets demonstrated its scalability and adaptability to complex real-world networks.

In conclusion, this framework offers an effective and scalable solution for analyzing large-scale networks, making it a valuable tool for advancing research in network science and real-world applications. Future work will focus on optimizing scalability, extending the framework to dynamic and heterogeneous networks, and automating parameter tuning.

References

- Ackoff, R. L. (1961). 'Management misinformation systems', *Management Science*, 14(4), pp. 147–156.
- Akoglu, L. and Tong, H. (2023). 'Graph anomaly detection and description: A comprehensive survey', *IEEE Transactions on Knowledge and Data Engineering*, 35(3), pp. 712–735.
- Bonacich, P. (1987). 'Power and centrality: A family of measures', *American Journal of Sociology*, 92(5), pp. 1170–1182.
- Brandes, U. (2001). 'A faster algorithm for betweenness centrality', *Journal of Mathematical Sociology*, 25(2), pp. 163–177.
- Brin, S. and Page, L. (1998). 'The anatomy of a large-scale hypertextual web search engine', *Proceedings of the 7th International Conference on World Wide Web (WWW)*, Elsevier, pp. 107–117.
- Chen, Y., Zhang, Y., Wang, T., and Yu, P. S. (2021). 'Edge-enhanced graph neural networks for community detection', *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, pp. 1299–1307.
- Freeman, L. C. (1978). 'Centrality in social networks conceptual clarification', *Social Networks*, 1(3), pp. 215–239.
- Fortunato, S. and Hric, D. (2016). 'Community detection in graphs', *Nature Reviews Physics*, 1(4), pp. 215–230.
- Gao, C., Liang, Y., and Shi, Z. (2020). 'Exploiting attributed graphs for detecting anomalies in communities', *Knowledge-Based Systems*, 203, p. 106174.
- Gao, J., Wang, W., et al. (2022). 'Anomaly and community detection in attributed graphs: An integrated approach', *IEEE Transactions on Knowledge and Data Engineering*, 34(9), pp. 4037–4049.
- Liu, J., Yin, H., and Cai, S. (2022). 'Attributed graph anomaly detection with multiple graph attention networks', *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, ACM, pp. 1218–1227.
- Newman, M. E. J. (2006). 'Modularity and community structure in networks', *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577–8582.
- Newman, M. (2018). *Networks*. Oxford University Press.
- Rossetti, G. and Cazabet, R. (2018). 'Community discovery in dynamic networks: A survey', *ACM Computing Surveys*, 51(2), pp. 1–37.
- Singhal, A. (2001). 'Modern information retrieval: A brief overview', *IEEE Data Engineering Bulletin*, 24(4), pp. 35–43.
- Wu, L., Pan, S., Chen, F., et al. (2021). 'A comprehensive survey on graph neural networks', *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), pp. 4–24.
- Zhang, C., Yang, C., Sun, L., et al. (2021). 'Bayesian robust attributed graph clustering', *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), pp. 4518–4525.
- Zhang, X., Xu, C., and Wu, C. (2022). 'Towards graph neural networks in community detection: A survey', *ACM Computing Surveys*, 54(10), pp. 1–32. GIL.