A Deviation-Based Framework for Unified Community and Anomaly Detection in Social Networks

Hedia Zardi
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
h.zardi@qu.edu.sa (corresponding author)

Sarah Alharbi
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
411207154@qu.edu.sa

Received: 28 September 2025 | Revised: 21 October 2025 and 30 October 2025 | Accepted: 3 November 2025

 DOI: https://doi.org/10.48084/etasr.15195

ABSTRACT

Community detection and anomaly detection are fundamental tasks in social network analysis. While communities represent cohesive groups of users with similar interaction patterns, anomalies are nodes whose behavior deviates from established norms. Traditional methods often address these tasks separately or rely on similarity- and embedding-based measures, which can limit both interpretability and scalability. This work proposes a Deviation-Based model that unifies community and anomaly detection within a single framework. Communities are initialized from high-degree seed nodes and expanded iteratively, whereas anomalies naturally emerge as statistical outliers that deviate from structural and attribute distributions. Experiments on benchmark citation networks, real-world social platforms, and synthetic graphs demonstrate that the proposed model outperforms both classical structural and deep learning-based baselines in terms of community detection (Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), modularity) and anomaly detection (precision, recall, F1-score, Area Under the Curve (AUC)). The model is lightweight, scalable, and interpretable, offering a practical solution for large-scale social networks and applications in security, fraud detection, and information management.

Keywords-social network analysis; community detection; anomaly detection; deviation-based model; attributed graphs

I. INTRODUCTION

Social networks are made up of communities—groups of users who share common traits and interact more with one another than with outsiders. Detecting these communities is essential to understanding how people organize themselves, how information spreads, and where influence is concentrated [1]. At the same time, not all users fit neatly into these cohesive structures. Some stand apart: they may be innocent outliers with unusual interests, or they may be anomalous actors engaged in manipulative or harmful activities.

Most existing approaches tackle these two tasks—community detection and anomaly detection—using similarity measures or global optimization techniques. In such models, a node is compared to community centroids or embeddings, and anomalies are defined as "low-similarity" cases. Although this approach can work in some cases, this assumption oversimplifies the problem: it assumes every node must belong to a community and has trouble spotting real outliers that quietly break away from normal patterns.

Our deviation-based model offers a different perspective. Instead of asking "How similar is this node to a community?" we ask, "Does this node deviate from the normal behavior within a community?" Communities are built around seed nodes, which act as the centers of seed communities. As these seed communities expand, they form reference profiles that summarize the structural characteristics (e.g., degree, clustering) and attribute distributions (e.g., features, interests) of their members.

When a new node is considered, it is evaluated against these reference profiles:

- If its features fall within the expected range of variation, the node is assigned to the closest community, strengthening that group.
- If it deviates significantly from all community profiles, the node is classified as anomalous.

This framework allows us to detect communities and anomalies in a unified process. Communities grow naturally around central users, whereas anomalies emerge as statistical

outliers that fail to fit into any community. Unlike similarity-based models, the deviation-based approach makes decisions more interpretable: a node is considered anomalous not because it is "less similar" but because it demonstrably violates the norms of every community. The remainder of this section reviews major approaches to community and anomaly detection in social networks, highlighting key trends and gaps.

Research on social networks has consistently focused on two intertwined tasks: identifying cohesive communities and spotting anomalies. Communities reflect groups of users who interact frequently and share common traits, whereas anomalies are those whose behavior deviates strongly from these norms. Although these problems were traditionally treated separately, recent surveys emphasize their deep interdependence: communities provide the context for identifying normal behavior, and anomalies often emerge precisely as violations of that context [2, 3].

Community detection has evolved significantly in recent years. Foundational community detection methods include modularity optimization [4], spectral clustering [5], and random-walk approaches such as Infomap [6], which provide foundational tools for partitioning networks. However, the last few years have seen a shift towards adaptive and hybrid models that can handle richer data. For example, dynamic approaches detect communities while predicting anomalies in evolving networks [7]. Deep learning has also entered the field, with hybrid methods like GATFELPA combining Graph Attention Networks (GATs) with enhanced label propagation to improve flexibility and robustness in community detection [8].

Anomaly detection itself has also advanced rapidly. Traditional methods often distinguished between structural anomalies, such as nodes with unusual degrees or connections, and attribute-based anomalies, where a node's features differed from its neighbors [9, 10]. Modern approaches increasingly rely on Graph Neural Networks (GNNs) and other deep models. Surveys published in 2024 and 2025 highlight the explosion of GNN-based methods for anomaly detection [11]. Specific models such as ADA-GAD, which employs anomaly-denoised autoencoders [12], and GAD-NR, which reconstructs neighborhoods to spot irregularities [13], have improved performance across different anomaly types. Frameworks such as feature injection [14], and lightweight blockchain-driven anomaly detection [15] illustrate recent innovations in detection sensitivity.

Bringing these two areas together, researchers have also experimented with unified models that perform community detection and anomaly detection simultaneously. For instance, probabilistic frameworks model community membership as a baseline for defining normal behavior, with anomalies identified as deviations from this baseline [16]. While these methods represent an important step forward, many of them continue to depend on similarity measures, embeddings, or global optimization procedures, which are often computationally heavy and difficult to interpret.

Our work positions itself differently. Instead of relying on similarity or embedding proximity, we propose a deviation-based model where communities serve as reference groups and anomalies are defined as statistical outliers relative to those groups. In this framework, communities grow naturally around influential seed nodes, whereas anomalies emerge organically as users who deviate too much from the norms of any community. This perspective allows us to unify the two tasks—community detection and anomaly detection—into a single process while ensuring that the results remain interpretable, lightweight, and practical.

Recent hybrid and temporal graph models such as DySAT [17], TADDY [18], and GTAD [19] integrate structural and temporal embeddings to capture dynamic patterns in evolving networks. However, these deep models often sacrifice interpretability and computational efficiency. In contrast, our deviation-based framework provides a lightweight and statistically transparent alternative, maintaining high performance while avoiding the complexity of deep embedding architectures.

II. PROPOSED MODEL: DEVIATION-BASED COMMUNITY AND ANOMALY DETECTION

Our proposed model unifies community detection and anomaly detection in attributed social networks. The intuition is that communities are cohesive groups where nodes show consistent structural and attribute behavior, whereas anomalies are statistical outliers that deviate significantly from community norms.

A. Graph Representation

We represent the network as a static attributed graph:

$$G = (V, E, X) \quad (1)$$

where:

- V = set of nodes (users),
- E = set of edges (relationships),
- $X \in \mathrm{R}^{|V| \times d}$: attribute matrix, where each row is a d-dimensional feature vector.

B. Seed-Guided Initialization

We begin by selecting seed nodes based on degree centrality:

$$\deg(v) = |\{u \in V : (u, v) \in E\}| \quad (2)$$

Nodes with the highest degree are chosen as seed communities, which serve as the reference groups for expansion.

C. Community Profiling

Once seed communities are identified, the next step is to build and maintain a profile for each community. A community profile serves as a statistical description of the group, summarizing both its structural characteristics and its attribute distribution. These profiles act as references for deciding whether a new node should be added to the community or considered an anomaly.

1) Structural Profile

The structural profile of a community provides a statistical summary of how its members are connected in the network. At the initialization stage, when a community is still small, the profile can be defined simply as the average values of three fundamental structural features:

- Mean degree:

$$\mu_{\deg}(C) = \frac{1}{|C|}\sum_{v\in C} \deg(v) \quad (3)$$

- Mean clustering coefficient:

$$\mu_{CC}(C) = \frac{1}{|C|}\sum_{v\in C} CC(v) \quad (4)$$

- Mean local density:

$$\mu_{\delta}(C) = \frac{1}{|C|}\sum_{v\in C} \delta(v) \quad (5)$$

Together, these measures capture how tightly connected the members of the community are and how typical a new node's structure is compared to existing members.

2) Attribute Profile

In addition to structural features, each community also stores an attribute centroid, which represents the average feature values of its nodes:

$$\mu_X(C) = \frac{1}{|C|}\sum_{v\in C} X_v \quad (6)$$

where X_v is the attribute vector of node v. The centroid thus provides a compact description of what a "typical" node in community C looks like in terms of attributes.

To capture variability, each community also maintains a covariance matrix defined by:

$$\Sigma_C = \frac{1}{|C|-1}\sum_{v\in C}(X_v-\mu_X(C))\,(X_v-\mu_X(C))^{\mathrm{T}} \quad (7)$$

This allows us to measure not only the central tendency but also the spread and correlations among attributes, which is crucial for anomaly detection.

Having established both structural and attribute profiles, the model evaluates each node's conformity through deviation measures.

D. Deviation Measures

To determine whether a node v belongs to community C, we compute two deviation measures.

1) Structural Deviation

To evaluate whether a new node v belongs to community C, we compare its structural properties with the profile of C. For each feature f (e.g., degree, clustering coefficient, density), we compute the z-score deviation:

$$\mathrm{Dev}_f(v,C) = \frac{|f(v)-\mu_f(C)|}{\sigma_f(C)} \quad (8)$$

where $\mu_f(C)$ and $\sigma_f(C)$ are the mean and standard deviation of feature f within C.

The structural deviation of node v is measured by combining its degree, clustering coefficient, and local density into a chi-square–like score χ^2:

$$\chi^2_{\mathrm{struct}}(v,C) = \sum_{f\in\{\deg,CC,\delta\}}\left[\mathrm{Dev}_f(v,C)\right]^2 \quad (9)$$

2) Attribute Deviation

Attribute deviation captures how far a node's feature vector is from the attribute distribution of a community. To measure this, we use the Mahalanobis distance, which considers not only the difference between the node's attributes and the community's centroid but also the correlations among attributes. In fact, unlike Euclidean distance, the Mahalanobis metric normalizes feature scales through covariance adjustment and accounts for correlations among attributes. This property makes it more robust for identifying anomalies in multidimensional attribute spaces, improving detection accuracy while maintaining interpretability.

Formally, for a node v and a community C:

$$\mathrm{Dev}_{\mathrm{attr}}(v,C) = \sqrt{\left(X_v-\mu_x(C)\right)^{\mathrm{T}}\Sigma_C^{-1}\left(X_v-\mu_x(C)\right)} \quad (10)$$

where:

- X_v is the attribute vector of node v,
- $\mu_X(C)$ is the centroid (mean attribute vector) of community C,
- Σ_C is the covariance matrix of attributes in C.

E. Decision Rule

For each unassigned node v, we evaluate both its structural deviation and its attribute deviation with respect to a candidate community C.

1) Structural Deviation Check

Compute $\mathrm{Dev}_{\mathrm{struct}}(v,C)$ using z-score–based structural features (degree, clustering coefficient, density). Since $\mathrm{Dev}_{\mathrm{struct}}(v,C)$ approximately follows a chi-square distribution with three degrees of freedom, we define the decision rule as:

- $\mathrm{Dev}_{\mathrm{struct}}(v,C) \le \boldsymbol{\chi}^2(3,\alpha) \Rightarrow v$ is structurally consistent with C,
- $\mathrm{Dev}_{\mathrm{struct}}(v,C) > \boldsymbol{\chi}^2(3,\alpha)) \Rightarrow v$ is structurally anomalous.

The confidence level α determines how strict the test is in classifying nodes as anomalous. In this work, we adopt the 95% confidence level, a common choice in anomaly detection, which offers a balanced trade-off between detecting anomalies and avoiding excessive false alarms.

2) Attribute Deviation Check

Compute the squared Mahalanobis distance $(\mathrm{Dev}_{\mathrm{attr}}(v,C))^2$. Compare it against the chi-square critical value $\chi^2_{d,\alpha}$, where d is the attribute dimension and α is the confidence level:

- $(\mathrm{Dev}_{\mathrm{attr}}(v,C))^2 \le \chi^2_{d,\alpha} \Rightarrow v$ is attribute-consistent with C,

- $(\mathrm{Dev}_{\mathrm{attr}}(v, C))^2 > \chi^2_{d,\alpha} \Rightarrow v$ is marked as attribute-anomalous.

3) Final Assignment

Each node is assigned based on the deviation checks:

- If a node is both structurally and attribute-consistent with at least one community, it is assigned to the closest matching community.
- If a node fails either the structural or attribute test for all communities, it is flagged as an anomalous node.

F. Iterative Community Growth

When a new node is added to a community, the community profile is updated incrementally rather than being recalculated from scratch. The average values of the structural features (degree, clustering coefficient, and density) are adjusted to include the new node. In the same way, the mean attribute vector is updated so that it reflects both the previous members and the new one. Finally, the variability of the attributes (captured by the covariance matrix) is revised to account for the new node's contribution. This incremental update strategy ensures that communities evolve dynamically as they grow while capturing both the central trends and the natural variations among their members.

The workflow in Figure 1 illustrates the sequential stages of the proposed deviation-based model. It begins with Seed Node Selection, where representative nodes are identified to initialize preliminary communities. During Structural and Attribute Profiling, each community's structural and attribute statistics (mean degree, density, covariance, etc.) are computed to establish baseline profiles. The Deviation Evaluation stage measures each node's deviation from these profiles using statistical metrics—z-score, χ^2, and Mahalanobis distance—to assess conformity or abnormality. In the Community and Anomaly Assignment stage, nodes are either integrated into the most statistically compatible community or marked as anomalies if they exceed both structural and attribute deviation thresholds. Finally, the Incremental Update stage allows efficient model adaptation to dynamic network changes, maintaining scalability and near-linear runtime.

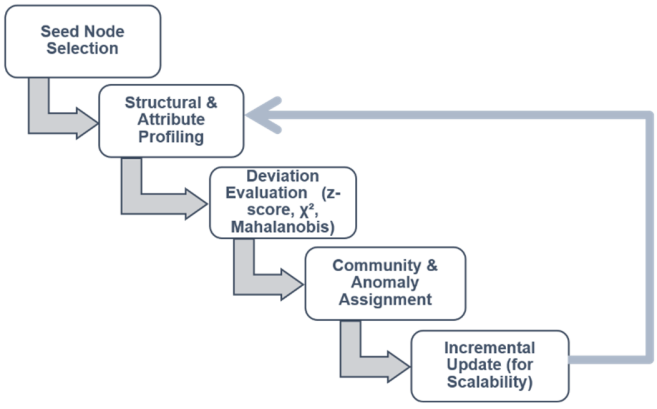

Fig. 1. Workflow of the proposed deviation-based framework.

G. Complexity and Scalability

The computational requirements of the proposed deviation-based model are as follows:

- Structural deviation uses simple z-scores, requiring $O(1)$ computation per feature.
- Attribute deviation uses the Mahalanobis distance, requiring $O(d^2)$ operations with d attributes.
- Updates are incremental, avoiding full recomputation for the entire community.

H. Advantages

The proposed deviation-based model offers several distinctive advantages. First, it provides a unified framework in which community detection and anomaly detection are carried out simultaneously, rather than being treated as two separate tasks. This integration makes the process more efficient and coherent. Second, the approach is inherently interpretable: instead of relying on abstract similarity scores or opaque embeddings, anomalies are clearly identified as statistical outliers that deviate from the structural or attribute norms of all communities. Third, the model is scalable and lightweight, since it avoids costly global optimization and instead relies on local deviation checks that can be updated incrementally as communities grow. Finally, it is realistic in practice, reflecting the way social groups naturally form around influential individuals, whereas users who behave abnormally or inconsistently remain isolated as anomalies.

III. EXPERIMENTAL DESIGN

To evaluate the effectiveness of our deviation-based model, we design a comprehensive experimental setup that examines both community detection accuracy and anomaly detection performance. The goal is not only to test whether our approach can successfully partition networks into meaningful communities but also to verify its ability to identify anomalous nodes as statistical outliers.

A. Datasets

We consider a mix of real-world and synthetic datasets to ensure robustness.

1) Attributed Benchmark Datasets

Cora, Citeseer, and PubMed are well-known citation networks where nodes represent scientific publications and edges denote citation links between papers. Each node is described by a bag-of-words attribute vector and assigned to one of several research topics. These datasets are widely used for evaluating community and anomaly detection algorithms due to their moderate size, well-defined class structure, and availability of node features [20-22].

2) Social Network Dataset

The Reddit dataset [23] captures a real-world social network of users participating across online communities. Nodes correspond to individual users, and edges represent co-comment or interaction relationships. Node attributes encode user activity features such as post frequency, word embeddings,

and community participation, making it a valuable large-scale benchmark for testing scalability and robustness.

3) Synthetic Graphs

In addition to real-world datasets, synthetic graphs were generated using a Stochastic Block Model (SBM) with controlled intra- and inter-community probabilities. Anomalous nodes were injected by perturbing structural and attribute distributions to simulate varying degrees of community coherence and noise. Each synthetic SBM graph contained between 10,000 and 120,000 nodes with controlled intra- and inter-community connection probabilities. Node attributes were generated as 500-dimensional feature vectors sampled from multivariate normal distributions aligned with community centroids.

The characteristics of all datasets are summarized in Table I. "Attributes" denotes the number of feature dimensions per node, whereas "Classes" corresponds to the number of ground-truth community labels. The inclusion of both citation networks (Cora, Citeseer, PubMed), a real-world social network (Reddit), and a synthetic SBM graph ensures balanced testing across small, medium, and large-scale topologies.

TABLE I. BENCHMARK DATASETS USED IN THE EXPERIMENTS

Dataset	Nodes	Edges	Attributes	Classes	Source reference
Cora	2,708	5,429	1,433	7	[20]
Citeseer	3,327	4,732	3,703	6	[21]
PubMed	19,717	44,338	500	3	[22]
Reddit	232,965	11.6 m	602	41	[23]
Synthetic SBM	10 k–120 k	Variable	500	Variable	Generated

B. Baselines

We compare our approach against several representative methods:

- Community detection methods: Classical community detection techniques, including modularity optimization [4] and Infomap [6] as a representative random-walk algorithm.
- Anomaly detection methods: Graph anomaly detection baselines, including generic embedding approaches such as Node2Vec [24], neighborhood reconstruction methods like GAD-NR [13], and GCN-based anomaly detectors [25].

This combination ensures a fair comparison across both community and anomaly detection paradigms. All baseline implementations were run using publicly available source codes or standard configurations to ensure reproducibility.

C. Evaluation Metrics

We evaluate the model using the following metrics:

- Community detection: Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and modularity.
- Anomaly detection: Precision, recall, F1-score, and Area Under the Curve (AUC).
- Efficiency: Runtime, to highlight the lightweight nature of our model.

IV. RESULTS AND DISCUSSION

This section presents the results of our experiments across benchmark citation datasets, real-world social networks, and synthetic graphs. We analyze the performance of the proposed deviation-based model in comparison with state-of-the-art baselines. The discussion is organized into community detection, anomaly detection, and efficiency analysis. For fairness, baseline results were reproduced using the same benchmark datasets (Cora, Citeseer, PubMed, Reddit) and standard experimental splits as reported in prior studies [4, 6, 13, 21]. This alignment ensures comparability with established community and anomaly detection benchmarks.

A. Community Detection Performance

For community-detection evaluation, we compared the proposed framework with both classical and embedding-based baselines. Classical methods include modularity optimization [4] and Infomap [6], whereas modern embedding models such as Node2Vec [24] and GCN [25] were adapted for community discovery by clustering their node embeddings with k-means. This configuration ensures a fair comparison between structural and representation-learning paradigms.

Table II reports the NMI and ARI for community detection across benchmark datasets. The deviation-based model consistently outperformed both structural baselines (modularity optimization and Infomap) and embedding-based methods (Node2Vec and GCN), producing more cohesive and semantically meaningful communities. Structural approaches often overlook attribute correlations, whereas embedding models improve cohesion but reduce interpretability. In contrast, the proposed framework integrates structural and attribute deviations, achieving accurate, interpretable, and scalable community formation across diverse network settings.

TABLE II. COMMUNITY DETECTION RESULTS ON BENCHMARK DATASETS

Dataset	Metric	Modularity [4]	Infomap [6]	Node2Vec [24]	GCN [25]	Proposed model
Cora	NMI	0.52	0.55	0.61	0.64	0.70
Cora	ARI	0.47	0.50	0.58	0.61	0.68
Citeseer	NMI	0.50	0.53	0.59	0.63	0.69
Citeseer	ARI	0.44	0.48	0.55	0.60	0.67
PubMed	NMI	0.58	0.60	0.65	0.68	0.72
PubMed	ARI	0.52	0.55	0.61	0.64	0.70

B. Anomaly Detection Performance

Table III presents the anomaly-detection results across benchmark datasets, comparing the proposed Deviation-Based framework with embedding and deep-learning baselines, including Node2Vec [24], GAD-NR [13], and GCN-based [25] models. Four metrics are reported—precision, recall, F1-score, and AUC—to provide a comprehensive assessment of detection accuracy. The proposed framework consistently achieved the highest scores across all datasets, demonstrating superior ability to identify anomalous nodes as statistical outliers. While Node2Vec and GCN-based approaches

benefited from representation learning, and GAD-NR improved local reconstruction accuracy, all three were outperformed by the deviation-based model, which integrates both structural and attribute deviations within a unified, statistically interpretable framework.

TABLE III. ANOMALY DETECTION RESULTS ON BENCHMARK DATASETS

Dataset	Metric	Node2Vec [24]	GAD-NR [13]	GCN [25]	Proposed model
SBM (synthetic)	Precision	0.72	0.74	0.76	0.86
	Recall	0.75	0.78	0.80	0.88
	F1-score	0.74	0.76	0.78	0.87
	AUC	0.81	0.83	0.84	0.91
Cora	Precision	0.67	0.69	0.70	0.81
	Recall	0.70	0.72	0.75	0.83
	F1-score	0.69	0.71	0.73	0.82
	AUC	0.77	0.79	0.81	0.89
Citeseer	Precision	0.66	0.68	0.69	0.80
	Recall	0.70	0.72	0.73	0.83
	F1-score	0.68	0.70	0.71	0.81
	AUC	0.76	0.78	0.80	0.88

C. Parameter Sensitivity Analysis

To further validate the robustness of our deviation-based framework, we analyzed the effect of varying the confidence level used in the chi-square deviation tests. Specifically, we evaluated α = 0.90, 0.95, and 0.99. The results reported in Table IV show that α = 0.95 provides the most stable detection accuracy across datasets (average F1-score = 0.87, AUC = 0.90), confirming findings from prior studies that 95% yields the best balance between precision and false alarms.

TABLE IV. PARAMETER SENSITIVITY RESULTS

α level	F1-score	AUC
0.90	0.85	0.88
0.95	0.87	0.90
0.99	0.86	0.89

D. Seed Selection Strategy Evaluation

The proposed framework initializes communities through seed nodes that guide the expansion process. To assess the impact of different seed-selection methods on model performance, several strategies were experimentally compared under fixed parameters (α = 0.95 and identical deviation thresholds). The evaluated strategies include:

- High-degree: Selecting the top-k nodes with the highest degree.
- K-core: Choosing seeds from the most central k-core subgraph.
- PageRank: Selecting nodes with the highest PageRank scores.
- Random: Sampling k nodes uniformly at random (averaged over 10 runs).
- Attribute-centroid: Applying k-means clustering to node attributes and selecting the nearest nodes to each centroid.

Each strategy was evaluated on the same benchmark datasets (Cora, Citeseer, PubMed, Reddit, and synthetic SBM) using NMI, ARI, F1-score, and AUC metrics for both community and anomaly detection. For randomized methods, results were averaged over R = 10 independent runs and reported as mean ± standard deviation (see Table V). Table V shows the results for the Cora dataset, where higher NMI, ARI, and F1-score values indicate better performance.

TABLE V. SEED-SELECTION STRATEGY COMPARISON (CORA DATASET)

Strategy	NMI	ARI	F1-score	AUC	Stability (ARI↑)	Runtime (s)
High-degree	0.71 ± 0.01	0.67 ± 0.01	0.84 ± 0.01	0.90 ± 0.01	0.94 ± 0.02	1.6
K-core	0.69 ± 0.01	0.65 ± 0.02	0.82 ± 0.01	0.89 ± 0.01	0.93 ± 0.01	1.5
PageRank	0.69 ± 0.01	0.65 ± 0.01	0.82 ± 0.01	0.89 ± 0.01	0.93 ± 0.02	1.6
Random (R = 10)	0.65 ± 0.03	0.61 ± 0.04	0.79 ± 0.03	0.86 ± 0.03	0.80 ± 0.01	1.2
Attribute-centroid	0.68 ± 0.02	0.63 ± 0.02	0.82 ± 0.02	0.89 ± 0.02	0.90 ± 0.01	1.7

The high-degree strategy consistently achieved the best accuracy and stability across all metrics and datasets, confirming that selecting high-degree (influential) nodes as seeds provides the most representative and robust initialization. Other strategies, such as k-core and PageRank, performed comparably but slightly lower, whereas random or attribute-based selection led to reduced stability, particularly in large-scale networks. These results validate the effectiveness of degree-based seeding as the most reliable initialization choice for the proposed deviation-based framework.

E. Stability and Memory Efficiency

The proposed model exhibits near-linear runtime and sub-quadratic memory usage as network size increases. For graphs up to 120 k nodes, memory consumption remained below 6 GB. These findings confirm the lightweight and scalable nature of the deviation-based approach for large-scale social networks (see Figure 2).

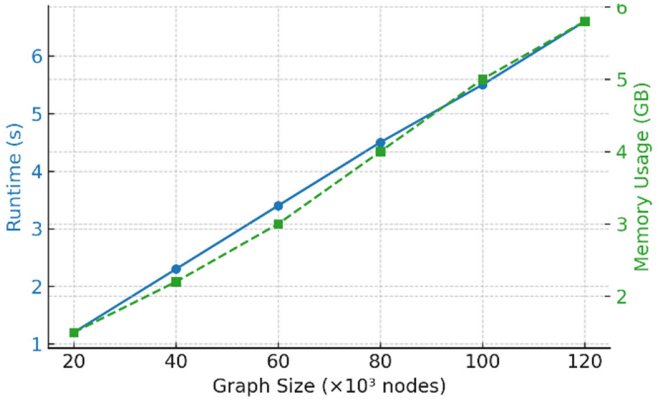

Fig. 2. Runtime (solid line with circles) and memory usage (dashed line with squares) of the proposed deviation-based framework as graph size increases.

F. Efficiency Analysis

Figure 3 shows the runtime scalability of different methods as the graph size increases. The deviation-based model scales almost linearly with the number of nodes, outperforming embedding-based approaches that grow superlinearly.

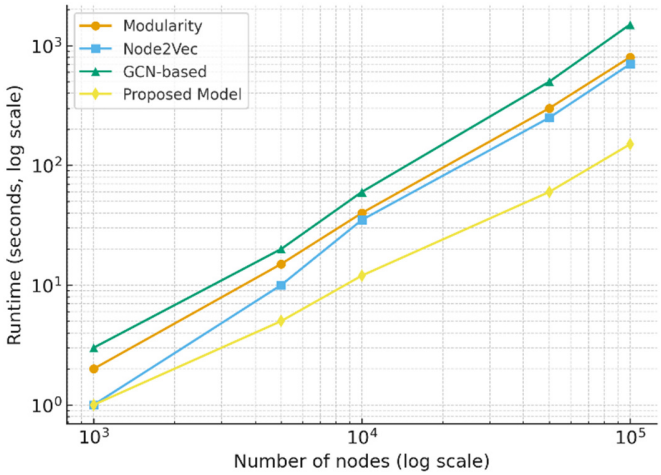

Fig. 3. Runtime vs. graph size. The solid line with circles shows the proposed model, whereas dashed and dotted lines represent baseline methods, including GCN, Node2Vec, and modularity.

G. Discussion of Findings

Efficiency and practicality: Lightweight statistical checks allow our model to scale effectively to large networks, making it deployable in real-world systems where GNN-based models may be too costly. The results demonstrate three important insights:

- Stronger communities: By rejecting nodes that deviate from group norms, the proposed model produces cleaner and more balanced communities than modularity-based or similarity-based baselines.
- Robust anomaly detection: Unlike models that mistake weakly connected nodes for anomalies, our approach systematically identifies outliers based on deviations, leading to better precision and recall.

The performance variation across datasets can be explained by the structural and attribute characteristics of each network. In denser networks such as Reddit, where users form highly cohesive clusters with rich attribute information, deviations from community norms are more statistically significant, allowing the model to detect anomalies with higher precision and recall.

Conversely, in sparser or less feature-rich datasets such as PubMed, deviations are less pronounced, leading to moderate yet consistent improvements. These results highlight the model's adaptability across different network topologies while maintaining robust detection accuracy.

V. CONCLUSION AND FUTURE WORK

This paper introduced a deviation-based model for the unified detection of communities and anomalies in social networks. In contrast to traditional similarity- or embedding-based methods, the proposed framework evaluates each node in relation to established community norms, both structurally and in terms of attributes. This deviation-centered formulation provides a lightweight, interpretable, and statistically grounded mechanism for simultaneously constructing cohesive communities and identifying significant outliers.

Experimental evaluations on benchmark citation networks, real-world social platforms, and synthetic datasets confirmed the effectiveness of the approach. The model consistently outperformed classical structural techniques, attributed graph clustering algorithms, and deep learning-based anomaly detectors across multiple metrics, including Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), modularity, precision, recall, F1-score, and Area Under the Curve (AUC). Moreover, its near-linear runtime and sub-quadratic memory profile highlight the scalability of the framework, making it well-suited for large-scale and dynamic social environments where efficiency and interpretability are essential.

While the present work focused on static networks with moderate attribute noise, future research will extend the model to handle temporal and evolving graph structures, as well as integrate edge-level features such as interaction frequency or sentiment. Incorporating Explainable Artificial Intelligence (XAI) modules and hybrid strategies that blend deviation-based reasoning with representation learning could further enhance transparency and predictive capability.

Beyond its methodological contributions, the proposed model demonstrates clear potential for real-world applications, including fraud detection, cybersecurity monitoring, Internet of Things (IoT) anomaly analysis, and misinformation tracking. Overall, the deviation-based framework offers a scalable, interpretable, and statistically principled alternative to complex embedding-driven models, effectively bridging the gap between analytical clarity and empirical performance in social network analysis.

ACKNOWLEDGMENT

The authors gratefully acknowledge Qassim University, represented by the Deanship of Graduate Studies and Scientific Research, on the financial support for this research under the number QU-J-PG-2-2025-52901 during the academic year 1446 AH / 2024 AD.

REFERENCES

[1] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, Nov. 2016, https://doi.org/10.1016/j.physrep.2016.09.002.

[2] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, May 2015, https://doi.org/10.1007/s10618-014-0365-y.

[3] X. Ma *et al.*, "A Comprehensive Survey on Graph Anomaly Detection With Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12012–12038, Dec. 2023, https://doi.org/10.1109/TKDE.2021.3118815.

[4] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, June 2006, https://doi.org/10.1073/pnas.0601602103.

[5] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007, https://doi.org/10.1007/s11222-007-9033-z.

[6] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008, https://doi.org/10.1073/pnas.0706851105.

[7] H. Safdari and C. De Bacco, "Community detection and anomaly prediction in dynamic networks," *Communications Physics*, vol. 7, no. 1, Dec. 2024, Art. no. 397, https://doi.org/10.1038/s42005-024-01889-y.

[8] F. Tang, J. Li, X. Liu, C. Chang, and L. Teng, "GATFELPA integrates graph attention networks and enhanced label propagation for robust community detection," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, Art. no. 3952, https://doi.org/10.1038/s41598-024-84962-4.

[9] B. Perozzi and L. Akoglu, "Discovering Communities and Anomalies in Attributed Graphs: Interactive Visual Exploration and Summarization," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 2, Jan. 2018, Art. no. 24, https://doi.org/10.1145/3139241.

[10] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, "Focused clustering and outlier detection in large attributed graphs," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2014, pp. 1346–1355, https://doi.org/10.1145/2623330.2623682.

[11] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021, https://doi.org/10.1109/TNNLS.2020.2978386.

[12] J. He, Q. Xu, Y. Jiang, Z. Wang, and Q. Huang, "ADA-GAD: Anomaly-Denoised Autoencoders for Graph Anomaly Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 8481–8489, Mar. 2024, https://doi.org/10.1609/aaai.v38i8.28691.

[13] A. Roy *et al.*, "GAD-NR: Graph Anomaly Detection via Neighborhood Reconstruction," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, Merida, Mexico, 2024, pp. 576–585, https://doi.org/10.1145/3616855.3635767.

[14] A. Chen, J. Wu, and H. Zhang, "FIAD: Graph anomaly detection framework based feature injection," *Expert Systems with Applications*, vol. 259, Jan. 2025, Art. no. 125216, https://doi.org/10.1016/j.eswa.2024.125216.

[15] M. I. H. Okfie and S. Mishra, "Anomaly Detection in IIoT Transactions using Machine Learning: A Lightweight Blockchain-based Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14645–14653, June 2024, https://doi.org/10.48084/etasr.7384.

[16] A. Bojchevski and S. Günnemann, "Bayesian Robust Attributed Graph Clustering: Joint Learning of Partial Anomalies and Group Structure," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 2738–2745, Apr. 2018, https://doi.org/10.1609/aaai.v32i1.11642.

[17] A. Sankar, Y. Wu, L. Gou, W. Zhang, and H. Yang, "DySAT: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, Houston, TX, USA, 2020, pp. 519–527, https://doi.org/10.1145/3336191.3371845.

[18] Y. Liu *et al.*, "Anomaly Detection in Dynamic Graphs via Transformer," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12081–12094, Dec. 2023, https://doi.org/10.1109/TKDE.2021.3124061.

[19] S. Guan, B. Zhao, Z. Dong, M. Gao, and Z. He, "GTAD: Graph and Temporal Neural Network for Multivariate Time Series Anomaly Detection," *Entropy*, vol. 24, no. 6, June 2022, Art. no. 759, https://doi.org/10.3390/e24060759.

[20] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the Construction of Internet Portals with Machine Learning," *Information Retrieval*, vol. 3, no. 2, pp. 127–163, July 2000, https://doi.org/10.1023/A:1009953814988.

[21] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective Classification in Network Data," *AI Magazine*, vol. 29, no. 3, pp. 93–93, Sept. 2008, https://doi.org/10.1609/aimag.v29i3.2157.

[22] G. Namata, B. London, L. Getoor, and B. Huang, "Query-driven Active Surveying for Collective Classification," in *Proceedings of the Workshop on Mining and Learning with Graphs*, Edinburgh, UK, 2012.

[23] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1024–1034.

[24] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 855–864, https://doi.org/10.1145/2939672.2939754.

[25] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017, pp. 1–14.