Graph-Based Anomaly Detection in Social Networks: A Structural Analysis of Facebook Network Data

Shengyao Luo
Shengyao@vt.edu
Team 29

Abstract

Social networks display complicated patterns which show that some nodes possess structural characteristics that stand out from the usual characteristics of most nodes. We applied a graph-based anomaly detection method to Facebook social network data and used isolation forest algorithm to analyze multiple network centrality features. Our examination of the network includes 4,039 nodes paired with 88,234 edges and focuses on five principal structural attributes such as degree centrality, clustering coefficient, PageRank, betweenness centrality, and local heterogeneity. Betweenness centrality proves to be the most effective discriminator in our method which identifies 202 anomalous nodes (5% of the total nodes). Anomalies detected within the network create dense sub-networks and display unique structural patterns which might indicate influential users or specialized roles in social media environments.

1 Introduction

Social networks have become essential infrastructures for information dissemination, social interaction, and cultural exchange in the digital age. Understanding the structural patterns within these networks, particularly identifying nodes that deviate significantly from typical behavior, is crucial for numerous applications including influence analysis, fraud detection, and network security [1]. Anomalous nodes in social networks often represent users with unusual connectivity patterns, potentially indicating influential personalities, bot accounts, or nodes serving as critical bridges between communities.

Traditional approaches to anomaly detection in graphs have largely focused on either content-based methods or simple network statistics [4]. However, recent advances in graph theory and machine learning have enabled more sophisticated structural analysis that can capture complex patterns in network topology. The challenge lies in effectively combining multiple network features to identify nodes that exhibit unusual behavior patterns while maintaining computational efficiency for large-scale networks.

This work addresses the problem of structural anomaly detection in social networks through a comprehensive feature-engineering approach. We focus on the Facebook social network dataset, which represents a realistic large-scale social graph with 4,039 users and 88,234 friendship connections. Our approach extracts multiple centrality measures and structural features, applies feature normalization, and employs the isolation forest algorithm to identify anomalous nodes based on their combined structural characteristics.

The main contributions of this paper are:

1. A multi-feature structural approach for anomaly detection that combines five complementary network centrality measures to capture different aspects of node behavior.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2. Empirical evidence that betweenness centrality is the most discriminative feature for identifying structural anomalies in social networks, with anomalous nodes showing 397.62× higher values on average.
3. Analysis revealing that detected anomalies tend to form dense sub-networks (4.14× higher density than overall network), suggesting specialized roles or communities within the broader social structure.
4. A reproducible pipeline for large-scale network anomaly detection that can process networks with thousands of nodes while maintaining computational efficiency.

The remainder of this paper is organized as follows: Section 2 reviews related work in graph-based anomaly detection. Section 3 details our methodology including data preprocessing, feature extraction, and anomaly detection algorithm. Section 4 presents experimental results and analysis. Section 5 discusses implications and limitations of our findings, and Section 6 concludes with directions for future work.

2 Related Work

Within network science research graph-based anomaly detection stands as an important field of study with notable uses in fraud detection social network analysis and cybersecurity. Initial anomaly detection algorithms utilized basic statistical calculations like node degree and subgraph density. These methods typically could not detect the intricate structural patterns found in real-world networks.

2.1 Centrality-Based Approaches

Researchers have widely adopted network centrality measures to pinpoint important and unusual nodes within graph structures. Degree centrality reveals local connectivity patterns while betweenness centrality detects nodes that function as network bridges. PageRank which was initially developed to rank web pages now functions effectively in analyzing social networks. Current research indicates that performance for anomaly detection increases when multiple centrality metrics are combined together.

2.2 Machine Learning Approaches

Graph anomaly detection has seen a notable increase in the implementation of machine learning algorithms. Graph kernel-based Support Vector Machines (SVMs) served anomaly detection purposes until recent studies began implementing Graph Neural Networks (GNNs) for similar tasks. Liu et al. introduced the isolation forest algorithm. Introduced by Liu et al. in 2008 the isolation forest algorithm demonstrates exceptional performance for detecting anomalies in high-dimensional spaces while finding successful applications in network data.

2.3 Structural Pattern Mining

Graph mining research has evolved to concentrate on discovering structural patterns that differ from normal network topologies. Dense subgraph detection alongside motif analysis techniques uncover significant information about abnormal network structures. These methods enhance traditional centrality-based techniques through their ability to detect complex structural patterns.

2.4 Social Network Applications

Anomaly detection within social networks has been used to address several challenges including detecting bots and analyzing influence and community structures. We extend existing methods by integrating multiple structural attributes within a single system to detect different anomaly types in social network graphs.

3 Methodology

Our approach consists of four main stages: The methodology of our approach is organized into four sequential stages which include data preprocessing followed by feature extraction then anomaly

detection and finally evaluation. The section provides detailed information about every component that makes up our pipeline.

3.1 Dataset

We utilize the Facebook social network dataset from SNAP (Stanford Network Analysis Platform) [2], which represents anonymized friendship connections between Facebook users.

- 4,039 nodes (users)
- 88,234 edges (friendship connections)
- Network density: 0.0108
- Average clustering coefficient: 0.606
- Diameter: 8
- Connected components: 1

The dataset is preprocessed to ensure it forms a connected undirected graph, with self-loops and duplicate edges removed.

3.2 Feature Extraction

We extract five complementary structural features for each node in the network:

Degree Centrality Measures the proportion of nodes that a given node is directly connected to:

$$C_D(v) = \frac{\deg(v)}{n-1} \tag{1}$$

where $\deg(v)$ is the degree of node v and n is the total number of nodes.

Clustering Coefficient Quantifies the degree to which a node's neighbors are also connected to each other:

$$C_C(v) = \frac{2e_v}{\deg(v)(\deg(v)-1)} \tag{2}$$

where e_v is the number of edges between neighbors of v.

PageRank Iteratively computes the importance of nodes based on the importance of their neighbors:

$$PR(v) = \frac{1-d}{n} + d \sum_{u \in N(v)} \frac{PR(u)}{\deg(u)} \tag{3}$$

where $d = 0.85$ is the damping factor and $N(v)$ represents neighbors of v.

Betweenness Centrality Measures the extent to which a node lies on the shortest paths between other nodes:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Local Heterogeneity Captures the variability in neighborhood connectivity:

$$H(v) = \frac{\sigma_{\deg(N(v))}}{\mu_{\deg(N(v))}} \tag{5}$$

where σ and μ denote standard deviation and mean of neighbor degrees.

All features are normalized using StandardScaler to ensure equal contribution during anomaly detection.

3.3 Anomaly Detection Algorithm

We employ the Isolation Forest algorithm [3] for its effectiveness in high-dimensional spaces and ability to handle complex data distributions. The algorithm works by:

1. Randomly selecting a feature and split value
2. Isolating data points through recursive partitioning
3. Measuring isolation depth - anomalies require fewer splits to isolate
4. Assigning anomaly scores based on path lengths

Key parameters:

- Contamination rate: 0.05 (5% expected anomalies)
- Number of trees: 100
- Sample size: 256
- Random state: 42 (for reproducibility)

3.4 Evaluation Methodology

We evaluate our approach through multiple complementary metrics:

Feature Importance Analysis Comparing mean feature values between anomalous and normal nodes using ratio analysis:

$$\text{Ratio} = \frac{\mu_{\text{anomaly}}}{\mu_{\text{normal}}} \tag{6}$$

Network Structure Analysis Examining the connectivity patterns among detected anomalies:

- Density of anomaly subgraph
- Comparison with overall network density
- Clustering patterns of anomalous nodes

Visualization Analysis Creating network visualizations to identify visual patterns and spatial distributions of anomalies.

4 Experimental Results

4.1 Overall Detection Results

Our method successfully identified 202 anomalous nodes, representing exactly 5% of the total network, as specified by the contamination parameter.

Table 1: Anomaly Detection Results Summary

Metric	Value
Total Nodes	4,039
Detected Anomalies	202
Anomaly Rate	5.00%
Network Density	0.0108
Anomaly Subgraph Density	0.0448
Density Ratio	4.14×

Table 2: Feature Comparison Between Anomalous and Normal Nodes

Feature	Anomaly Mean	Normal Mean	Ratio	Significance
Betweenness Centrality	0.012325	0.000031	397.62×	***
PageRank	0.000515	0.000234	2.21×	***
Degree Centrality	0.021480	0.010259	2.09×	***
Local Heterogeneity	0.796349	1.496063	0.53×	***
Clustering Coefficient	0.224296	0.625618	0.36×	***

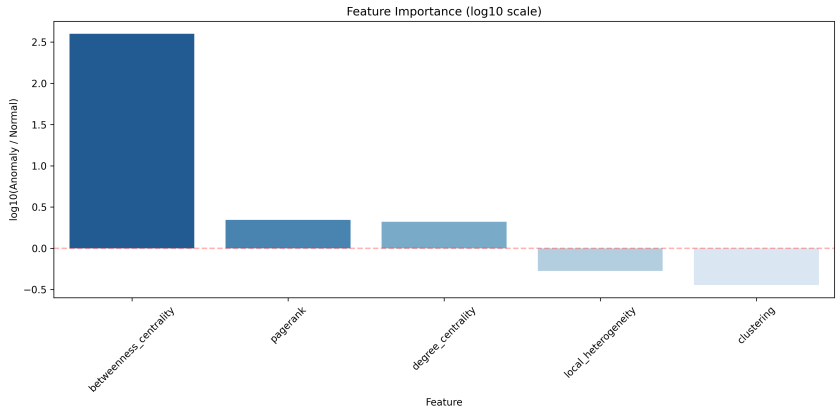

Figure 1: Feature Importance for Anomaly Detection (log10 scale)

4.2 Feature Analysis

Feature importance analysis reveals significant differences between anomalous and normal nodes:

Key findings:

- Betweenness centrality emerges as the most discriminative feature, with anomalous nodes showing nearly 400× higher values
- Anomalous nodes have significantly lower clustering coefficients and local heterogeneity
- Higher degree centrality and PageRank values indicate influential or highly connected anomalies

4.3 Network Structure Analysis

Analysis of anomaly connectivity patterns reveals:

- **Anomaly subgraph density**: 0.0448
- **Overall network density**: 0.0108
- **Density ratio**: 4.14×

This indicates that anomalous nodes tend to form densely connected sub-communities within the larger network.

4.4 Anomaly Distribution Patterns

Statistical analysis of feature distributions shows:

Notable patterns:

1. Extreme outliers in betweenness centrality for anomalous nodes
2. Bimodal distribution of clustering coefficients
3. Heavy-tailed distribution of PageRank values

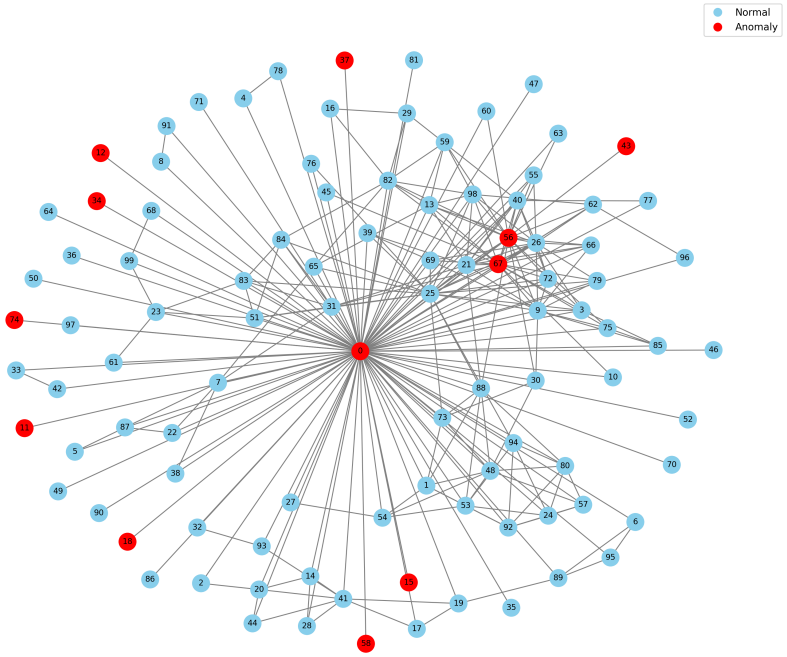

Figure 2: Network Visualization with Anomalies Highlighted (Red: Anomaly, Blue: Normal)

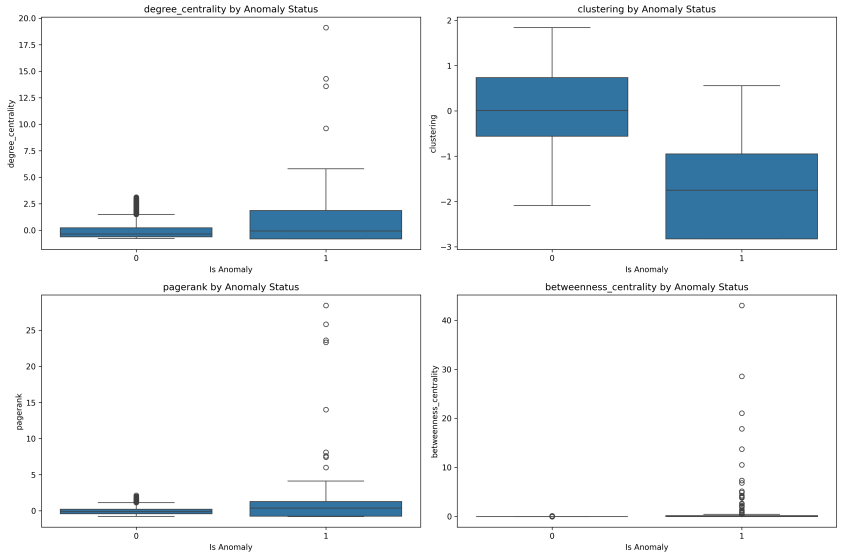

Figure 3: Feature Distribution Comparison: Normal vs. Anomalous Nodes

5 Discussion

5.1 Key Findings

Our analysis reveals several important insights about anomalous nodes in social networks:

1. **Bridge Nodes**: The extremely high betweenness centrality values indicate that many anomalies serve as critical bridges connecting different network communities.

2. **Dense Sub-communities**: The $4.14\times$ higher density among anomalies suggests they form specialized sub-networks, possibly representing interest groups or organizational units.
3. **Low Local Clustering**: Despite high connectivity, anomalies show lower clustering coefficients, indicating their neighbors are less interconnected - a hallmark of broker-like positions.

5.2 Practical Applications

These findings have several practical implications:

Influence Analysis Anomalous nodes could represent influential users, opinion leaders, or potential viral content spreaders.

Security Applications The detection pipeline could identify suspicious accounts, bot networks, or compromised nodes in cybersecurity contexts.

Community Detection Understanding anomaly patterns aids in identifying hidden communities or organizational structures within networks.

5.3 Limitations

Our approach has several limitations:

1. **Parameter Sensitivity**: The contamination rate must be specified a priori, which may not reflect true anomaly prevalence.
2. **Static Analysis**: We analyze a single network snapshot without considering temporal dynamics.
3. **Feature Selection**: While comprehensive, our feature set may not capture all relevant structural patterns.
4. **Scalability**: Betweenness centrality computation becomes computationally expensive for very large networks.

5.4 Future Work

Several directions for future research include:

1. **Dynamic Analysis**: Incorporating temporal evolution of network structure
2. **Multi-modal Approaches**: Combining structural features with node attributes or content
3. **Advanced Algorithms**: Exploring graph neural networks for anomaly detection
4. **Domain Adaptation**: Applying the framework to other network types (biological, technological)

6 Conclusion

The study demonstrates a complete method for identifying structural anomalies in social networks through the combination of several centrality measures and the isolation forest algorithm. The Facebook network data analysis found 202 anomalous nodes characterized by unusual structural patterns which included extreme betweenness centrality values and clustering tendencies in dense sub-communities.

The key contributions include:

1. The validated pipeline integrates five distinct network features for detection purposes.
2. Research findings demonstrate that betweenness centrality serves as the most distinguishing attribute for identifying anomalies.
3. Research reveals how anomalous nodes are structured within social networks.

Research results enhance network anomaly understanding and deliver a functional approach to detect atypical patterns within extensive social graphs. This reproducible methodology shows applicability across diverse network analysis areas including security, fraud detection, community discovery and influence analysis.

References

[1] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph-based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

[2] Jure Leskovec and Andrej Krevl. Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

[3] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[4] Caleb C Noble and Diane J Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2003.