Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Graph-Based Data Mining for Social Network Analysis

Vijay Kumar Meena
Lecturer,Govt. R.C Khaitan Polytechnic College,Jaipur
Email:-vijaysattawan22@gmail.com

ARTICLE INFO

Received: 05 March 2022

Accepted: 30 May 2022

ABSTRACT

Social networks generate massive amounts of relational data capturing interactions, influence, and community structures among users. Mining such data is crucial for understanding information diffusion, detecting communities, and identifying influential nodes. Traditional graph mining approaches often struggle with scalability and complex feature representations in large social networks. This paper proposes a graph-based data mining framework leveraging Graph Neural Networks (GNNs) for community detection and influence analysis. Our methods integrate node embeddings, graph convolutional layers, and attention mechanisms to capture structural and relational information. Experiments on large-scale Twitter datasets demonstrate that our approach achieves higher modularity scores for community detection and superior influence prediction accuracy compared to baseline methods such as Louvain clustering and PageRank. Quantitative metrics and visualizations illustrate the effectiveness of GNN-based techniques in identifying hidden communities and influential nodes, enabling actionable insights for social network analysis, marketing, and misinformation detection.

Keywords: Graph Neural Networks, Social Network Analysis, Community Detection, Influence Prediction, Node Embeddings, Twitter Graphs.

I. Introduction

Social networks such as Twitter, Facebook, and Instagram are platforms where **information spreads rapidly** through user interactions. Understanding network structures and influence patterns is critical for applications including viral marketing, epidemic modeling, misinformation detection, and recommendation systems [1].

Graph-based data mining provides a natural representation of social networks, where **nodes represent users and edges represent interactions** such as follows, likes, or retweets. Key tasks in social network analysis include:

1. **Community Detection:** Identifying groups of users with dense intra-group connections and sparse inter-group connections.
2. **Influence Analysis:** Determining influential nodes whose actions propagate widely across the network.

Conventional approaches for these tasks include **modularity-based clustering**, **centrality measures** (degree, betweenness, PageRank), and matrix factorization methods [2]. However, these methods often fail to capture **complex non-linear dependencies**, multi-hop relationships, and dynamic interaction patterns present in real-world social networks.

Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Graph Neural Networks (GNNs) have emerged as powerful tools for learning **node representations and edge relationships** by aggregating information from neighborhood nodes. GNNs can model high-order dependencies and leverage node attributes alongside graph topology [3].

This research focuses on:

- Developing **GNN-based algorithms** for community detection and influence prediction in large-scale social networks.
- Evaluating the framework on **Twitter graphs** with millions of nodes and edges.
- Measuring performance using **modularity, normalized mutual information (NMI), and influence prediction accuracy**.
- Comparing against baseline methods such as **Louvain clustering, Label Propagation, and PageRank**.

II. Related Work

A. Community Detection in Social Networks

Community detection seeks **densely connected subgraphs** representing user groups with shared interests or behaviors. Traditional methods include:

- **Modularity-based clustering:** Louvain algorithm maximizes modularity to identify communities [4].
- **Label Propagation Algorithm (LPA):** Propagates labels iteratively across neighbors to detect communities [5].
- **Spectral Clustering:** Uses eigenvectors of the graph Laplacian for partitioning [6].

Limitations:

- Sensitive to **initialization** and network size.
- Fail to capture **node attribute information**.
- Struggle with **overlapping communities**.

B. Influence Analysis

Identifying influential nodes is essential for **information diffusion modeling**. Traditional approaches include:

- **Centrality Measures:** Degree, betweenness, closeness, and eigenvector centralities [7].
- **PageRank:** Models random walks to identify nodes with high importance [8].
- **Independent Cascade Models:** Simulate influence spread [9].

Limitations:

- Do not leverage **node feature representations**.

Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- May ignore **higher-order interactions** in large-scale networks.

C. Graph Neural Networks

GNNs aggregate features from **node neighborhoods** using neural network layers:

- **Graph Convolutional Networks (GCN):** Layer-wise propagation for semi-supervised node classification [10].
- **Graph Attention Networks (GAT):** Learn attention weights for neighbors, enhancing representation learning [11].
- **GraphSAGE:** Sampling-based inductive method for large graphs [12].

GNNs have been applied for **node classification, link prediction, and graph classification**. Their ability to capture **local and global structure** makes them suitable for social network analysis tasks.

III. Methodology

A. Graph Representation

Let G=(V,E) denote a social network graph where:

- V = set of nodes (users)
- E = set of edges (interactions)
- $X \in R^{|V| \times d}$ = node feature matrix

Edges can be weighted based on interaction frequency (e.g., number of retweets). Node features include user metadata (number of followers, activity score, account age) and embeddings of textual content.

B. Community Detection Using GNNs

We propose a **GNN-based community detection framework**:

1. **Input Layer:** Node features X and adjacency matrix A.
2. **Graph Convolution Layers:** Aggregate neighbor information:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)})$$

   where $\tilde{A} = A + I$, $\tilde{D}$ is the degree matrix, $W^{(l)}$ is the learnable weight, and σ is the activation function.

3. **Community Assignment:** Soft clustering using **k-means** or **mixture models** on node embeddings from the final layer.

   Loss Function: Minimize reconstruction loss of adjacency matrix and cross-entropy for labeled nodes.

Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

C. Influence Prediction Using GNNs

- Use **node embeddings from GNN layers** as input to a **feedforward network** predicting influence scores.
- Supervision via **historical propagation data**: Retweet cascades or message spread.
- Loss function: Mean squared error (MSE) for influence score regression.

D. Attention Mechanism

- Integrate **graph attention layers (GAT)** to learn neighbor importance weights.
- Attention coefficient between nodes i and j:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i||Wh_j]))}{\sum_{k\in\mathcal{N}_i}\exp(\text{LeakyReLU}(a^T[Wh_i||Wh_k]))}$$

where || denotes concatenation, and aaa is a learnable vector.

IV. Experimental Setup

A. Datasets

1. **Twitter Social Graph:** Nodes = 2M users, Edges = 8M interactions (retweets, mentions).
2. **Facebook Page-Page Network:** Nodes = pages, edges = interactions.
3. **Synthetic Benchmark Graphs:** For algorithm validation and scalability testing.

Node features include: follower counts, activity metrics, textual embeddings from posts.

B. Baseline Methods

- **Louvain Modularity Maximization**
- **Label Propagation Algorithm (LPA)**
- **PageRank for Influence**

C. Evaluation Metrics

Community Detection:

- **Modularity (Q):** Measures intra-community density relative to random graph.
- **Normalized Mutual Information (NMI):** Compares predicted and ground-truth communities.

Influence Prediction:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**

Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Top-K Accuracy:** Fraction of correctly identified top influential nodes.

Computational Efficiency:

- Runtime and memory usage for large-scale networks.

V. Results

A. Community Detection

Table I: Modularity Scores Comparison

Method	Twitter Graph	Facebook Graph	Synthetic Graph
Louvain	0.62	0.58	0.65
LPA	0.55	0.52	0.58
GCN + k-means	0.71	0.67	0.74
GAT + k-means	0.74	0.70	0.76

Observation: GNN-based methods achieve **higher modularity**, indicating improved community detection.

Table II: NMI Scores Comparison

Method	Twitter Graph	Facebook Graph
Louvain	0.61	0.57
LPA	0.54	0.52
GCN + k-means	0.69	0.66
GAT + k-means	0.72	0.69

B. Influence Prediction

Table III: Influence Prediction Performance

Method	MAE	RMSE	Top-100 Accuracy
Degree Centrality	0.128	0.184	0.62
PageRank	0.115	0.172	0.67
GCN Regression	0.092	0.138	0.78
GAT Regression	0.088	0.132	0.81

Observation: GNN-based models outperform traditional centrality measures and PageRank in predicting influential nodes.

Journal of Information Systems Engineering and Management
2022, 7(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

C. Scalability Analysis

- **Training time for GAT on 2M-node graph:** 3.2 hours (single GPU)
- **Memory footprint:** 24GB GPU memory
- GraphSAGE variant reduces training time to 1.8 hours using neighborhood sampling.

VI. Discussion

1. **Community Detection:** GNN embeddings capture **structural and feature information**, enabling better detection of overlapping or hidden communities compared to modularity-only methods.
2. **Influence Analysis:** Attention-based aggregation improves identification of influential nodes, accounting for **neighbor importance**.
3. **Trade-offs:**
    - GAT slightly more accurate than GCN but requires **higher computation**.
    - Neighborhood sampling (GraphSAGE) balances **efficiency and accuracy**.
4. **Practical Implications:**
    - Marketing: Identify influential users for targeted campaigns.
    - Misinformation: Detect early propagators of fake news.
    - Community management: Discover hidden groups for moderation or engagement.
5. **Limitations:**
    - Dynamic networks require **temporal graph models**.
    - Large-scale deployment may need **distributed graph processing frameworks**.

VII. Conclusion

This paper presents a **graph-based data mining framework using GNNs** for social network analysis. Our contributions include:

1. Development of **GCN and GAT-based methods** for community detection and influence prediction.
2. Extensive evaluation on **Twitter, Facebook, and synthetic graphs**, demonstrating superior performance over traditional methods in modularity, NMI, and top-K influence prediction.
3. Practical insights for **marketing, information diffusion, and network analysis** in large-scale social media platforms.

Future work includes:

- Incorporating **temporal graph neural networks** for dynamic social networks.
- Scaling GNNs via **distributed training frameworks** for billion-edge graphs.
- Exploring **multi-modal embeddings** (text, images, interactions) for richer representation.
- Investigating **fairness and bias mitigation** in influence prediction and community detection.

References

[1] M. Newman, *Networks: An Introduction*, Oxford University Press, 2010.

[2] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75–174, 2010.

[3] W. Hamilton, Z. Ying, J. Leskovec, "Inductive representation learning on large graphs," *NeurIPS*, 2017.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[5] U. N. Raghavan, R. Albert, S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, 2007.

[6] J. Shi, J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[7] L. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, 1978.

[8] S. Brin, L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, 1998.

[9] D. Kempe, J. Kleinberg, É. Tardos, "Maximizing the spread of influence through a social network," *KDD*, 2003.

[10] T. Kipf, M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.

[11] P. Velickovic, et al., "Graph attention networks," *ICLR*, 2018.

[12] W. Hamilton, et al., "Inductive representation learning on large graphs," *NeurIPS*, 2017.