Project Proposal for COMP 6321: Machine Learning

Parsa Kamalipour
Department of Computer Science
Concordia University
parsa.kamalipour@mail.concordia.ca

Abdulmoumen Al-Atrash
Department of Computer Science
Concordia University
a_alatr@live.concordia.ca

Aniket Roy
Department of Computer Science
Concordia University
aniket.roy@mail.concordia.ca

Abstract

Community detection, which is a main problem in social network analysis, looks to find clusters in graphs where nodes are more connected with each other than with the rest of the network. This is important in different fields like social network analysis and biology, and it helps in understanding complex systems and their organization. Moreover, the recent growth of social networks through the inherent interconnectedness of the internet has made properly analyzing these intricate structures increasingly important. Traditional methods like modularity optimization and spectral clustering have limitations in handling large networks and often miss detailed community structures. For instance, modularity-based methods rely too heavily on maximizing modularity, which often leads to missing small but significant communities within the network, especially when community sizes vary. One notable model, Graph Neural Networks (GNNs), has emerged as a powerful deep learning approach to solving the community detection problem. This project aims to solve these limitations by using machine learning, specifically GNNs, due to its strong capabilities of identifying key patterns in complex data networks by effectively learning the high-dimensional feature representations of nodes and communities. With these tools, the project tries to make community detection more accurate and scalable. This plan aligns with course objectives, where we apply theoretical knowledge to practical problems in network analysis and machine learning. It addresses gaps in the existing methodologies for solving community detection by highlighting the merit of GNNs over traditional methods.

1 Problem Statement

1.1 Introduction to Community Detection

Community detection is all about identifying clusters or groups within a network - which are called communities - where nodes are more densely connected to each other than to the rest of the network. This problem is very crucial in various domains, including social network analysis, biology, and information retrieval, as it helps uncover the underlying structure and functional organization of complex systems. Namely, studying Social network graph structures. Traditional methods, such as modularity optimization and spectral clustering, have been widely used for this purpose. However, these approaches often face challenges in handling large-scale networks and capturing intricate community structures, which is becoming more and more important every day, since Social Networks are becoming a part of everyone's life these days.

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

1.2 Machine Learning Context

Recent advancements and novelties in machine learning, particularly deep learning, have introduced lots of new methodologies for community detection. Graph Neural Networks (GNNs) have emerged as very powerful tools capable of learning complex patterns in graph-structured data. For instance, the Contrastive Deep Nonnegative Matrix Factorization (CDNMF) model integrates contrastive learning with deep nonnegative matrix factorization to enhance and improve on community detection by capturing both network topology and node attributes at the same time [8]. Additionally, the integration of community detection algorithms with GNNs has been shown to improve link prediction tasks in scientific literature networks, demonstrating the synergistic potential of combining these approaches [9].

1.3 Usage of Machine Learning in Community Detection

In this project, we aim to use machine learning techniques to address the challenges of community detection in large-scale networks, to be more precise in Social Networks. By using models such as GNNs and incorporating methods like optimization on previous ideas, we seek to develop an algorithm that can effectively identify communities with high accuracy and scalability. This approach aligns with the course objectives by applying machine learning methodologies to solve complex problems, thereby by doing this research we both learn a new Machine Learning methodology and also enhance our understanding of community structures in various real-world applications.

2 Motivation

This section presents a literature review to examine pre-existing knowledge in the area of the NP-Hard problem of 'Community Detection' in social networks. We examine relevant existing work to ensure that our proposed approach is justified based on the problem context and that it differs from already proposed solutions to the NP-Hard problem.

2.1 Optimization Methods

A rather popular approach to solving challenging and inherently complex machine learning problems such as community detection is optimization. The Learning-Based Genetic Algorithm (LGA) [1] [Identifying Communities in Complex Networks Using Learning-Based Genetic Algorithm] is designed to tackle the NP-hard nature of community detection problems. This method carefully combines Learning Automata with genetic operations to get around problems that genetic algorithms often have, such as unwanted premature convergence and local optima. The LGA achieves a notable improvement in accuracy of 26.47% on real-world networks and a 48.32% improvement on synthetic networks. Akbar et al. [2] propose a quantum mechanics-inspired optimization approach that aims to detect underlying complex patterns, specifically in ecological communities. This paper simulates quantum principles for optimization detections for identifying patterns in biodiversity changes due to environmental factors such as land use and climate change. MARLCD [3] is a multi-agent reinforcement learning algorithm designed for community detection within complex networks. The algorithm employs "agents" to independently explore communities within the network, updating actions on successful detections. MARLCD outperforms state-of-the-art methods such as GA-Net and Meme-Net when tested on both real and synthetic datasets. However, these three approaches lack generalizability on networks of varying domains and complexity, as their underlying assumptions make them effective within their respective applications but restrict their flexibility and may be computationally excessive when applied to networks with different structural patterns and complexities.

2.2 Modularity-Based Methods

Modularity-based algorithms are simpler and easily interpretable techniques that are widely known for their community detection ability by maximizing modularity. Modularity is a novel metric based on the density of connected nodes within communities compared to their density with the rest of the network. Chen et al. [5] introduce a hierarchical clustering algorithm that is applied to optimize Max-Min Modularity. The authors argue that traditional statistical inference procedures make strong assumptions that instances are independent, and that relation-based methods cannot distinguish between features of the network domain, both of which lead to problematic conclusions about the data and lower performance. The Hybrid Genetic Tabu (HGT) [4] method aims to solve the community detection problem by maximizing modularity through a combination of Genetic Algorithms (GA) and Tabu Search (TS). This hybrid approach achieved higher detection accuracy compared to methods such as the Louvain and Label Propagation and demonstrated a stronger inclination toward medium-sized networks. Nevertheless, modularity-based approaches face several limitations; most prominent is their strong reliance on maximizing modularity as the primary objective, which is known to often be unable to distinguish differences in community sizes and connectivity, leading to suboptimal detection accuracy.

2.3 Deep Learning Methods

When it comes to utilizing powerful machine learning models to handle complex data processing, feature extraction, and pattern recognition, deep learning most often immediately comes to mind. Liu et al. [10] discuss various deep learning techniques for community detection, including deep neural networks, graph neural networks, and deep graph embeddings. The paper highlights the unique ability of deep learning models such as GNNs to overcome limitations of heavily relying on adjacency matrices and node attribute matrices by encoding higher-dimensional feature representations of nodes and communities. These limitations, circumvented by appropriate deep learning techniques, are prominent in traditional community detection methods, which often lead to the loss of complex structural relationships. The deep transitive encoder [12] is a novel approach that transforms the network's adjacency matrix to capture indirect node relationships that traditional methods often miss. The autoencoder utilizes unsupervised transfer learning to effectively extract low-dimensional features from the network structure. This led to improved accuracy over traditional methods; however, this approach's reliance on the adjacency matrix transformation was found to be very computationally demanding, which poses a problem for large-scale real-world networks. Nooribakhsh et al. [11]. systematically provide a comprehensive overview of machine learning trends in tackling the community detection problem. The paper highlights the recent growth of deep learning applications on large-scale, complex networks due to their consistency in outperforming simpler methods, such as game-theoretic and clustering algorithms. Moreover, the authors emphasize its superiority over more traditional approaches, like density-based methods, which often lack the ability to grasp structural and feature-based characteristics of nodes in networks.

2.4 Conclusion

The literature review highlights the diverse and sophisticated attempts to solve the community detection problem in social networks, which is notoriously an NP-hard problem. To support our motivation, we highlight the advantages and limitations of several nuanced approaches in recent literature. Optimization-based methods like the Learning-Based Genetic Algorithm (LGA) and Quantum-Inspired Optimization produce significant accuracy improvements over other state-of-the-art approaches but face complications in generalizing when presented with networks of varying domains and complexities. Traditional methods, such as modularity-based algorithms, provide simplicity and interpretability that come at the cost of a strict reliance on maximizing modularity. This restriction introduces limitations in identifying community structures and an inability to adapt to diverse community networks. Deep learning models, especially graph neural networks (GNNs), offer powerful solutions that serve to overcome these limitations by utilizing high-dimensional feature representations and avoiding the heavy reliance on adjacency matrices present in other models. Although techniques such as the deep transitive encoder and deep graph embeddings can achieve remarkable improvements over community detection accuracy, they often come with substantial computational demands to run.

This literature review serves as a justification of the proposed approach and highlights the potential of GNNs in solving the NP-hard problem of community detection while simultaneously minimizing limitations that accompany existing machine learning methods.

3 Proposed Approach

The ever-increasing size of networks raises the need for scalable methods for community detection. Traditional methods like spectral clustering and modularity optimization though effective for smaller graphs, struggle with the scalability required for modern applications. To address this gap without sacrificing accuracy we propose to develop a scalable Graph Neural Network (GNN) model leveraging efficient graph processing techniques.

3.1 Enhancing GNN Scalability

3.1.1 Graph Sampling Techniques

We aim to explore the integration of graph sampling techniques such as node sampling, layer sampling and subgraph sampling into the GNN architecture. These methods by processing only a subset of nodes and their neighbors can help reduce computational overhead.

3.1.2 Partitioning Strategies

We will be investigating the use of graph partitioning algorithms which will enable dividing the network in smaller subgraphs. This approach will thereby enable parallel processing of the subgraphs, reducing the memory requirements for training the GNN.

3.1.3 Efficient Graph Convolutions

We will also be exploring the incorporation of optimized graph convolutional operations to limit the spread of the convolutional process to a fixed-size neighborhood. Inspired by innovations in fast localized spectral filtering [7], this bounded approach might be beneficial in preventing the exponential growths of computational costs.

3.2 Training and Evaluation

Following an approach similar to [13] and [6] we will initially train and evaluate the proposed model using synthetic labeled data from Stochastic Block Models (SBM). This phase will allow us to asses the model's baseline metrics. Through varying SBM parameters to simulate different levels of graph sparsity and noise, the model will be able to learn complex community patterns, as demonstrated in [13] and [6].

Subsequently, the model will be evaluated on real-world datasets from SNAP. This phase will be crucial in testing the model's performance in large-scale networks where community structures are less defined and there is often presence of irregular patterns and noise.

4 Conclusion

In conclusion, this project looks to overcome the limitations present in traditional community detection approaches by using advanced machine learning methods, specifically Graph Neural Networks (GNNs). Traditional methods, such as modularity-based and optimization methods, often struggle with generalizability and the ability to adapt to networks of varying sizes and structures. By focusing on both accuracy and scalability, this approach is aimed at making community detection more useful for large, complex networks.

This project not only applies theoretical ideas from the course but also shows how they can be used for real-world challenges in network analysis. Through this work, there's a chance to understand network structures better and how advanced machine learning models help improve community detection. This paper showcases the potential of advanced deep learning models, such as GNNs, to optimize the solutions to the NP-Hard problem of 'Community Detection' in social networks.

References

[1] Gholam Reza Abdi, Amir Hosein Refahi Sheikhani, Sohrab Kordrostami, Bagher Zarei, and Mohsen Falah Rad. Identifying communities in complex networks using learning-based genetic algorithm. *Ain Shams Engineering Journal*, page 103031, 2024.

[2] S. Akbar and S. K. Saritha. Quantum inspired community detection for analysis of biodiversity change driven by land-use conversion and climate change. *Scientific Reports*, 11(1):14332, 2021.

[3] Mir Mohammad Alipour and Mohsen and Abdolhosseinzadeh. A multiagent reinforcement learning algorithm to solve the community detection problem. *Signal and Data Processing*, 19(1), 2022.

[4] Bouchema Sara Cheikh Salmi and Zaoui Sara. An enhanced evolutionary approach for solving the community detection problem. *Journal of Information and Telecommunication*, 6(1):83–100, 2022.

[5] Jiyang Chen, Osmar R. Zaïane, and Randy Goebel. *Detecting Communities in Social Networks using Max-Min Modularity*, pages 978–989.

[6] Zhengdao Chen, Xiang Li, and Joan Bruna. Supervised community detection with line graph neural networks. *arXiv preprint arXiv:1705.08415*, 2017.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[8] Yuecheng Li, Jialong Chen, Chuan Chen, Lei Yang, and Zibin Zheng. Contrastive deep nonnegative matrix factorization for community detection, 2024.

[9] Chunjiang Liu, Yikun Han, Haiyun Xu, Shihan Yang, Kaidi Wang, and Yongye Su. A community detection and graph neural network based link prediction approach for scientific literature, 2024.

[10] Fanzhen Liu, Shan Xue, Jia Wu, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Jian Yang, and Philip S. Yu. Deep learning for community detection: progress, challenges and opportunities. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[11] Mahsa Nooribakhsh, Marta Fernández-Diego, Fernando González-Ladrón-De-Guevara, and Mahdi Mollamotalebi. Community detection in social networks using machine learning: a systematic mapping study. *Knowledge and Information Systems*, 66(12):7205–7259, December 2024.

[12] Ying Xie, Xinmei Wang, Dan Jiang, and Rongbin Xu. High-performance community detection in social networks using a deep transitive autoencoder. *Information Sciences*, 493:75–90, 2019.

[13] Shunjie Yuan, Chao Wang, Qi Jiang, and Jianfeng Ma. Community detection with graph neural network using markov stability. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 437–442. IEEE, 2022.