ICT 3229 Data Mining and Predictive Analysis
FISAC Assignment - 3 Minute Thesis Report

Paper Title - LiGNN: Graph Neural Networks at LinkedIn

Pragya Gupta - 220953230

9th April, 2025

1 Introduction

LinkedIn is a professional social networking website with over 1 billion users globally. LinkedIn's user base consists of members, companies, universities, students and more, who use this platform to find career opportunities and connect with each other, forming connections that are represented as a graph. Integrating these varied entities, is a complex challenge, as it contains hundred billion nodes and several hundred billion edges. Graph nodes symbolize the platform's users, while the edges describe the interactions between the users, such as job applications, post engagements, and networking. It is essential for LinkedIn to improve recommendation systems across various domains as it will help garner more users and increase their satisfaction.

2 Motivation and Problem Statement

A Graph Neural Network (GNN) is a deep learning model designed to operate on graph data structure. GNNs learn to collect information from each node and its neighbors through a series of message-passing steps, producing embeddings that capture local structure and feature information. By leveraging the power of neural networks with the relational structure of graphs, GNNs are integral to social network analysis. Traditional recommendation systems at LinkedIn relied on linear and ensemble tree models, which failed to capture complex feature interactions, and separated feature generation from modeling. This made it hard to leverage the relationship data that defines its network.

These limitations motivated the development of Graph Neural Networks at LinkedIn (LiGNN), which faced technical challenges such as: the scale of the generated graph, diverse data types, infrequent visits by some users resulting in less data (cold start), dynamic and temporal nature of the network, and training stability and efficiency. It is a fundamental engineering challenge to manage these along with memory constraints and computational complexity, and requires sophisticated modeling approaches.

These problems could not be solved trivially as they required simultaneous advances in algorithmic design, system architecture, and deployment strategies. The research question became: "How can GNNs be effectively scaled and optimized to enhance recommendation quality across LinkedIn while addressing the challenges of its network?"

3 Proposed Solution

LiGNN is a comprehensive framework for deploying GNNs. The solution encompasses innovative components addressing the challenges met.

3.1 GNN Modeling

1. **Graph Construction:** The graph has many types of nodes and edges. To combine them, subgraphs from different domains are merged, and train their GNNs using its subgraph, or the combined graph. Two-hop Personalized PageRank (PPR) sampling is used to capture the structural information, leading to more relevant recommendations.

2. **GNN Architecture:** An encoder-decoder architecture is employed for the GNN models. The encoder uses a GraphSAGE-style framework for learning, generating node embeddings through neighborhood aggregation, enabling generalization to unseen nodes. The decoder uses these embeddings as input to make predictions. Temporal encoder enhancements and long-term losses significantly improve predictive performance and adapts to evolving user behaviors and preferences.

3. **Graph Densification:** Most nodes have very few interactions, which is a challenge for neighborhood aggregation. Artificial edges are added to low-degree nodes based on similarity with neighboring nodes, which are found using KNN. This helps to solve cold start issues.

3.2 Training

1. **Stability:** GNNs demand real-time sampling from the Graph Engine (GE) and retrieval of labeled data, increasing network strain and reducing training stability. This is balanced by employing gRPC retry (+15%), Horovod training (+35%) and fixing memory leak issue (+10%).

2. **Speed:** During development, the training time decreased from 24 to 3.3 hours. This is done by:

   - Reducing average step time: Local gradient is accumulated to reduce communication overhead in data parallel setups (+35.2%) along with mixed precision in forward/backward passes (+8%).
   - Increase convergence speed: Node encoder weights are trained from features without querying the GE (+16.25%).
   - Adaptive Neighbor Sampling: Number of samples during training is adaptively increased by monitoring model performance, mitigating I/O bottleneck (+24.2%).
   - Grouping and Slicing: Records are grouped by member, slicing each interactions into fixed-size chunks, and querying each member's graph neighborhood only once per chunk, at the cost of slightly larger queries (+69.9%).
   - Python Multi-Processing with Shared Memory Queue: This is implemented across multiple processes to speed up data prefetching and preprocessing (+68.02%).

4 Evaluation and Results

LiGNN is evaluated through comprehensive A/B testing. Experimentation conducted:

1. **Follow Feed:** Recommendations were treated as a link prediction problem to estimate member-post interactions. +0.5% in feed engaged daily active users and +9.6% in recall metrics is observed.

2. **Out-Of-Network Feed:** Posts beyond immediate connections are recommended by extending GNNs. Online A/B tests showed +0.2% in daily active users.

3. **Job Recommendations:** When integrated into LinkedIn's Top Applicant Jobs feature, it improved position suggestions, +0.3% in premium member subscription renewal, +1% in +ve application response rates, and +1.8% in company follows, demonstrating effectiveness in connecting members with relevant job opportunities.

4. **People Recommendations:** +0.1% in new member connections and weekly active users.

5. **Ads:** Graph topology integrated into Click-Through Rate (CTR) prediction addresses sparsity resulting +2% in online CTR.

While the evaluation demonstrates significant improvements, further exploration can be done in explainability of the model for end-users and its long-term adaptability. The improvements across diverse use cases validates its generalizability. The evaluation results are impressive given the scale at which these were achieved, as +0.5% in daily active users represents millions of additional engaged users across its massive platform.

LiGNN relies on infrastructure using Kubernetes clusters, GPU nodes for training, and real-time GE. To make it accessible, detailed deployment guides could help other organizations adopt similar frameworks. Cloud-based solutions offering pre-configured pipelines could lower barriers.

5 Contribution

LiGNN contributes to GNNs and industry applications of machine learning such as scalable GNN framework and training efficiency. Valuable insights into system architecture and technology choices are provided, that can guide other organizations in similar endeavors. Improvements in key business metrics for LinkedIn translates to enhanced user satisfaction, improved engagement, and increased revenue.

6 Future Direction

While LiGNN represents a noteable advancement, some directions for future research are:

1. **Further Scaling:** As LinkedIn's network continues to grow, better scaling and inference will be necessary. Exploring distributed training across larger clusters and compression techniques could increase speed.

2. **Advanced Graph Operations:** More complex graph operations can be explored to capture patterns not currently utilized.

3. **Cross-Domain Transfer Learning:** Transfer learning between different recommendation tasks, leveraging insights from one domain to improve another can be used to enhance overall performance.

4. **Explainability:** Improving the explainability of recommendations would build user trust and increase transparency.

7 Conclusion

LiGNN represents an advancement in the recommendation systems at LinkedIn, addressing the challenges of scaling GNNs to operate on massive and dynamic graphs. By introducing innovative techniques, it has delivered improvements across LinkedIn's core business domains, including job recommendations, feed personalization, and advertising. Its ability to reduce training time by 7x and integration into real-world applications demonstrate its practicality and scalability. The results from online A/B testing validate its impact, and growth on key metrics translate to significant business value for LinkedIn, enhancing user satisfaction and driving growth. As professional networks and graph-based systems continue to evolve, frameworks like LiGNN will play a critical role in delivering personalized experiences to users around the globe.

DMPA_3MT_Report.pdf

ORIGINALITY REPORT

1%	1%	0%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		