Link Prediction of Companies Relationships using Graph Neural Network

Mile Pelivanov, Miroslav Mirchev and Igor Mishkovski

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Rudjer Boshkovikj 16, P.O. 393, 1000 Skopje, North Macedonia
mile.pelivanov@student.finki.ukim.mk
{miroslav.mirchev,igor.mishkovski}@finki.ukim.mk

Abstract. Recognizing the strategic importance of anticipating connections between companies, this documentation delves into link prediction for company relationships, utilizing the Relato Business Graph and employing Graph Neural Networks (GNNs). The focus is on binary prediction to determine link presence and absence, along with multi-class link prediction within the same graph. By leveraging tailored neural networks, the study aims to unveil patterns and dependencies in corporate networks, offering insights into the intricacies of company relationships. Through empirical findings, it contributes to the application of GNNs in predictive analytics for corporate network analysis, but also underscores the practical significance of accurately predicting links for informed decision-making in dynamic business environments.

I. INTRODUCTION

Recognizing the critical need to forecast relationships among companies, this documentation begins on an exploration of link prediction within corporate networks, employing the Relato Business Graph and harnessing the capabilities of Graph Neural Networks (GNNs). In an era characterized by increased interconnectivity and data abundance, the significance of understanding and predicting complex relationships between businesses has become more pronounced than ever. As our world becomes progressively more interconnected, we find ourselves in a landscape where large volumes of data on inter-corporate interactions are accessible, paving the way for sophisticated analytical approaches.

This study addresses the challenges and opportunities associated with predicting links between companies. The primary focus lies on leveraging the Relato Business Graph, a comprehensive representation of corporate relationships, and deploying advanced Graph Neural Networks (GNNs) to uncover hidden patterns within this complex web of connections. GNNs are proven to be powerful tools for various deep learning tasks, and their application in link prediction for company relationships holds promise for enhancing our understanding of the dynamic nature of corporate networks.

Within this framework, the study specifically examines two key aspects of link prediction. The first involves binary prediction, a fundamental task of determining the presence or absence of links between nodes in the corporate graph. This aspect provides a foundational understanding of the relationships, enabling a binary classification that informs whether two companies are connected or not. The second aspect extends the analysis to multi-class link prediction within the same graph structure. This adds a layer of complexity, as it seeks to categorize the relationships into multiple classes, providing a more detailed perspective on the nature of connections between companies.

To facilitate these predictions, four important architectures of Graph Neural Networks have been applied: GraphSAGE, GCN, GAT, and SGC. Each architecture has its unique strengths and characteristics, contributing to the overall predictive capabilities of the model. All these architectures introduce an additional layer of sophistication by

incorporating extra information such as node feature and edge label. The initial layers of these architectures focus on learning vector representations of the nodes, while the final layer is dedicated to making predictions about the classes of links.

In addition to the application of these advanced neural networks, the study incorporates post-processing techniques to refine predictions. One such technique involves passing the embeddings learned from the Graph Neural Network through a multi-layered perceptron. Other methods are also applied to enhance the accuracy and reliability of the predictions.

The significance of this research extends beyond the domain of academic exploration. Accurate link prediction in corporate networks holds immense practical importance for informed decision-making. It provides stakeholders with valuable insights into market trends, competitive landscapes, and emerging opportunities. By leveraging the power of Graph Neural Networks and the comprehensive representation offered by the Relato Business Graph, this documentation aims to not only contribute empirical findings to the field but also offer practical methodologies for enhancing predictive analytics in the dynamic landscape of corporate relationships.

In conclusion, the exploration of link prediction within corporate networks using Graph Neural Networks and the Relato Business Graph represents a significant step toward a more detailed understanding of the complexities inherent in inter-corporate dynamics. Through the application of advanced techniques and comprehensive analyses, this documentation seeks to empower stakeholders with the tools and insights necessary to navigate the intricate web of corporate connections, ultimately leading to more informed and strategic decision-making in the ever-evolving business environment.

II. DATASET

The business graph dataset includes a thorough collection of 373,663 company links sourced from the web. This dataset captures the complex web of connections between various businesses, providing valuable insights into their interrelationships. The links within this dataset are categorized into distinct types, including "partnership" denoting one company listed on another's partnership page, "customer" indicating one company listed on another's example customer page, "competitor" representing co-bidders, "investment" signifying a company listed on a venture capitalist's website, and "supplier" as the inverse of the "customer" type. The dataset, initially collected using Mongo, underwent further processing with Titan and Gremlin, enabling the definition of multi-relational metrics. The dataset's richness and diversity facilitated sophisticated queries, allowing the extraction of entire networks surrounding a company's current customers, partners, and competitors, offering a multifaceted perspective for in-depth analysis.

Given the extensive volume of the entire graph, we have strategically narrowed our focus to analyze a representative subset comprising 2000 samples, enabling efficient computational processing without compromising crucial patterns within the dataset. In our exploration, we experimented with various combinations of node features and edge labels to optimize our model's performance. In the initial approach, node features are characterized by the degree of the node, while edge labels are binary (1-0), signifying the existence or absence of a link between nodes. Further experimentation involves maintaining the same node features while changing the edge labels. Edges are categorized based on the type of relationship, with unique values assigned (1 for partnership, 2 for customer, etc.), and 0 denoting no link. This detailed edge labeling approach provides a richer understanding of the specific relationships between nodes, offering a more detailed perspective on the network structure. A significant augmentation to our exploration involves using the power of a GPT model. In this configuration, node names serve as inputs to the GPT model, and the resulting outputs provide valuable information about the field in which each node operates. This extracted field information from the GPT model is then employed as node features. Edge labels are still on a 0-6 scale, indicate the relationship type between nodes, providing a detailed representation of the graph's structure. Ultimately, our study explores three unique combinations of node features and edge labels, offering different perspectives on interpreting the dataset. This thorough approach enables us to evaluate the performance of different feature-label configurations in capturing meaningful patterns and relationships within the graph, enhancing our understanding of the complex web of connections among companies in this business graph database.

Fig 1. Relato Graph with all nodes and all relationships between each of the nodes

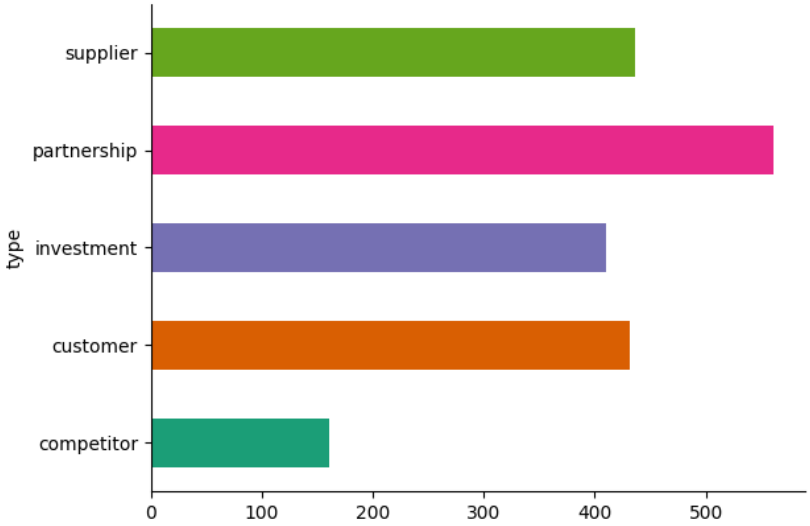

Fig 2. Distribution of each relationship in the dataset

Relationship	Number of relationships
Partnership	16876
Customer	3065
Investment	1423
Supplier	177
Competitor	719

Table 1. Number of relationships for each type

III. Binary Link Prediction

The Graph Neural Network models employed for the binary link prediction task include GraphSAGE, GCN, GAT, and SGC architectures. In the context of this task, each model is designed with a specific configuration, and they share a common objective: predicting the presence or absence of links between companies in the given graph. For instance, the GraphSAGE model is structured with two layers, transforming initial node embeddings into hidden representations and then into output predictions. The model utilizes SAGEConv layers and uses dropout for regularization. The first layer receives input vectors of size 2, representing the nodes, and the subsequent layers reduce these vectors dimensions to capture key features based on the graph structure. In simpler terms, the model learns weight matrices for both neighbor nodes and the node under consideration. By aggregating information from the neighborhood nodes and the given node, new embeddings are established. The final layer outputs predictions, and the model is trained using a binary cross-entropy loss function. During training, negative samples are generated, and the model aims to maximize the accuracy of true positive predictions while minimizing false positives. The GraphSAGE model is then optimized using the Adam optimizer.

Similarly, the GCN (Graph Convolutional Network) architecture is characterized by multiple layers that iteratively aggregate information from neighboring nodes. The initial node embeddings undergo transformations to capture hierarchical features based on the graph structure. The model excels at learning structural representations, contributing to accurate link predictions.

GAT (Graph Attention Network) introduces attention mechanisms to assign varying importance to neighbor nodes during aggregation. This allows the model to focus on relevant information, enhancing its ability to discern link patterns.

SGC (Simple Graph Convolution) simplifies the convolutional operation, making it computationally efficient while still capturing essential graph features. It offers a streamlined approach to link prediction. In summary, each model utilizes distinctive architectures to capture and interpret graph structures, enabling them to predict binary links effectively. The training process involves optimizing the models using the Adam optimizer and minimizing binary cross-entropy loss, while negative sampling helps in training the models to distinguish true and false positive predictions. The evaluation process assesses the model's performance on validation and test sets, providing insights into their predictive capabilities for binary links within the Relato Business Graph.

In our evaluation process, we employ the Open Graph Benchmark (OGB) link prediction evaluator. This evaluator is a valuable tool for assessing the performance of our binary link prediction models. To evaluate the effectiveness of our predictions, we utilize various metrics, including hits@20, hits@10, hits@5, hits@3, and hits@1. These metrics provide insights into the model's ability to correctly rank positive links among a set of candidate links. For instance, hits@20 measures the proportion of true positive links found within the top 20 ranked links, while hits@1 focuses on the accuracy of the top-ranked link. The evaluator evaluates our predictions based on the provided ground truth, allowing us to iteratively refine and optimize our link prediction models to achieve better results.

Table 2. Binary Link Prediction evaluated with Hits@20

Binary Link Prediction: Node Degree as Node Feature, 1-0 Link Existence as Edge Label, Evaluated with Hits@20				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0277	0.1365	0.2336	0.1622
Validation Hits	0.9875	0.7743	0.6050	0.8276
Test Hits	0.9548	0.5169	0.4053	0.6085
Binary Link Prediction: Node Degree as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@20				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0423	0.1711	0.1947	0.1495
Validation Hits	0.9655	0.6489	0.8088	0.8307
Test Hits	0.7604	0.5132	0.4003	0.5747
Binary Link Prediction: Industry Type as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@20				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0161	0.1653	0.2318	0.1480
Validation Hits	0.9937	0.7649	0.7900	0.7524
Test Hits	0.9649	0.5169	0.3764	0.5232

Table 3. Binary Link Prediction evaluated with Hits@10

Binary Link Prediction: Node Degree as Node Feature, 1-0 Link Existence as Edge Label, Evaluated with Hits@10				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0448	0.1702	0.2020	0.1604
Validation Hits	0.9624	0.6207	0.3636	0.5674
Test Hits	0.8984	0.3651	0.2535	0.4693
Binary Link Prediction: Node Degree as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@10				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0510	0.1992	0.2225	0.1570
Validation Hits	0.9498	0.6113	0.5204	0.7304
Test Hits	0.7980	0.2785	0.1593	0.5257
Binary Link Prediction: Industry Type as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@10				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0170	0.1749	0.2229	0.1499
Validation Hits	0.9718	0.5549	0.3793	0.6176
Test Hits	0.7691	0.3764	0.2208	0.4241

Table 4. Binary Link Prediction evaluated with Hits@5

Binary Link Prediction: Node Degree as Node Feature, 1-0 Link Existence as Edge Label, Evaluated with Hits@5				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0348	0.1859	0.2319	0.1777
Validation Hits	0.8370	0.5329	0.2696	0.5925
Test Hits	0.4040	0.2095	0.1104	0.2183
Binary Link Prediction: Node Degree as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@5				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0377	0.2031	0.2380	0.1561
Validation Hits	0.8182	0.5172	0.0658	0.5298
Test Hits	0.4743	0.1870	0.1405	0.2334
Binary Link Prediction: Industry Type as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@5				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0182	0.1610	0.1993	0.1866
Validation Hits	0.8370	0.3448	0.1630	0.5862
Test Hits	0.8269	0.2823	0.1217	0.1857

Table 5. Binary Link Prediction evaluated with Hits@3

Binary Link Prediction: Node Degree as Node Feature, 1-0 Link Existence as Edge Label, Evaluated with Hits@3				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0317	0.1698	0.2120	0.1577
Validation Hits	0.9154	0.4295	0.1411	0.4075
Test Hits	0.4567	0.1380	0.0464	0.1506
Binary Link Prediction: Node Degree as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@3				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0439	0.1679	0.2053	0.1504
Validation Hits	0.8809	0.4013	0.1661	0.4514
Test Hits	0.8168	0.0376	0.0640	0.2171
Binary Link Prediction: Industry Type as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@3				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0168	0.1485	0.2314	0.1748
Validation Hits	0.8621	0.5141	0.1944	0.4075
Test Hits	0.7516	0.3162	0.0427	0.1945

Table 6. Binary Link Prediction evaluated with Hits@1

Binary Link Prediction: Node Degree as Node Feature, 1-0 Link Existence as Edge Label, Evaluated with Hits@1				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0438	0.1537	0.2257	0.1672
Validation Hits	0.6928	0.3260	0.0188	0.2665
Test Hits	0.0000	0.0125	0.0125	0.0025
Binary Link Prediction: Node Degree as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@1				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0363	0.1630	0.2027	0.1564
Validation Hits	0.2226	0.2445	0.0909	0.4451
Test Hits	0.1192	0.0828	0.0125	0.0013
Binary Link Prediction: Industry Type as Node Feature, Type of Relationship as Edge Label, Evaluated with Hits@1				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.0157	0.1546	0.1994	0.1674
Validation Hits	0.5705	0.4326	0.0063	0.2288
Test Hits	0.1744	0.0100	0.0025	0.0276

IV. MULTICLASS LINK PREDICTION

For the multi-class link prediction task, we employed the same set of Graph Neural Network models, including GraphSAGE, GAT (Graph Attention Network), GCN (Graph Convolutional Network), and SGC (Simplified Graph Convolution, in conjunction with a Link Predictor. Their architecture consists of three GNN layers, with an input layer receiving a vector of size 2 representing the initial node embeddings. The second layer transforms these input vectors into vectors of size 32, leveraging the graph structure to learn weight matrices for both neighbor nodes and the node under consideration. By aggregating information from neighborhood nodes and the given node, new embeddings are established. The third and final layer is the output layer, with a dimension of the number of classes that are possible to be predicted. In this case, the models predict not only the presence or absence of links but also the type of link, using a softmax activation function.

The training process involves minimizing the cross-entropy loss between predicted and actual link types. The Neural Network models are optimized using the Adam optimizer. Their performance is evaluated using metrics such as F1 Score, Accuracy, Precision, and Recall on both validation and test sets, providing comprehensive insights into the model's ability to classify different link types within the graph.

During the process of multiclass link prediction, I'm using two approaches to enhance model understanding and prediction accuracy. Initially, I use node degree as a node feature, providing insights into the connectivity and importance of each node within the graph. However, recognizing that node degree may introduce biases towards highly connected nodes, I incorporate industry type as a node feature. This richer information contributes meaningful context to the node embeddings, offering a more detailed representation of each node's characteristics. Notably, in this multi-class link prediction scenario, where the goal is to classify links into various types, the edge labels, representing the types of links, remain unchanged throughout the process. This deliberate choice ensures that the model recognizes and predicts the diverse relationships existing in the graph without altering the fundamental nature of the link categories.

Table 7. Multiclass Link Prediction using Node Degree as Node Feature

Multiclass Link Prediction: Node Degree as Node Feature				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.2408	0.4482	0.5759	0.5508
Validation F1 Score	0.4818	0.3848	0.3052	0.3668
Validation Accuracy	0.3511	0.2806	0.2320	0.2665
Validation Precision	0.5013	0.4304	0.3308	0.3801
Validation Recall	0.4718	0.3750	0.3048	0.3671
Test F1 Score	0.4950	0.4057	0.3508	0.3598
Test Accuracy	0.3545	0.2917	0.2478	0.2622
Test Precision	0.5050	0.4076	0.3684	0.3659
Test Recall	0.4918	0.4069	0.3466	0.3681

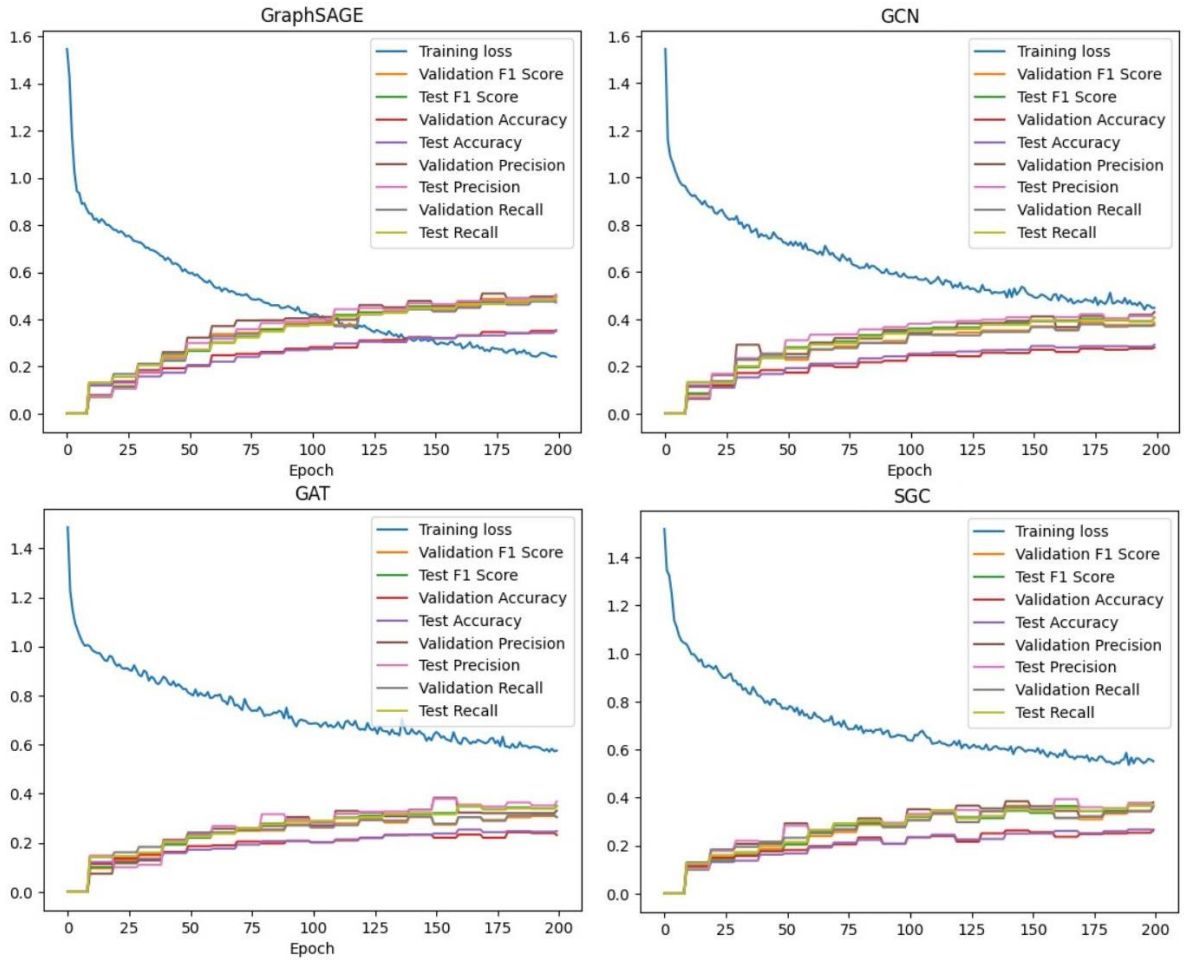

Fig 3. Training Loss and Model Performance on Validation and Test Sets over Epochs while using Node Degree as a Node Feature

Table 7. Multiclass Link Prediction using Industry Type as Node Feature

Multiclass Link Prediction: Node Degree as Node Feature				
	GraphSAGE	**GCN**	**GAT**	**SGC**
Loss	0.2506	0.4272	0.5427	0.4974
Validation F1 Score	0.5063	0.3943	0.3537	0.3950

Validation Accuracy	0.3636	0.2947	0.2665	0.2962
Validation Precision	0.5400	0.4761	0.4024	0.4000
Validation Recall	0.4927	0.3872	0.3446	0.3991
Test F1 Score	0.5069	0.4246	0.3728	0.4063
Test Accuracy	0.3614	0.3043	0.2691	0.2923
Test Precision	0.5159	0.4530	0.3916	0.4075
Test Recall	0.5028	0.4181	0.3685	0.4129

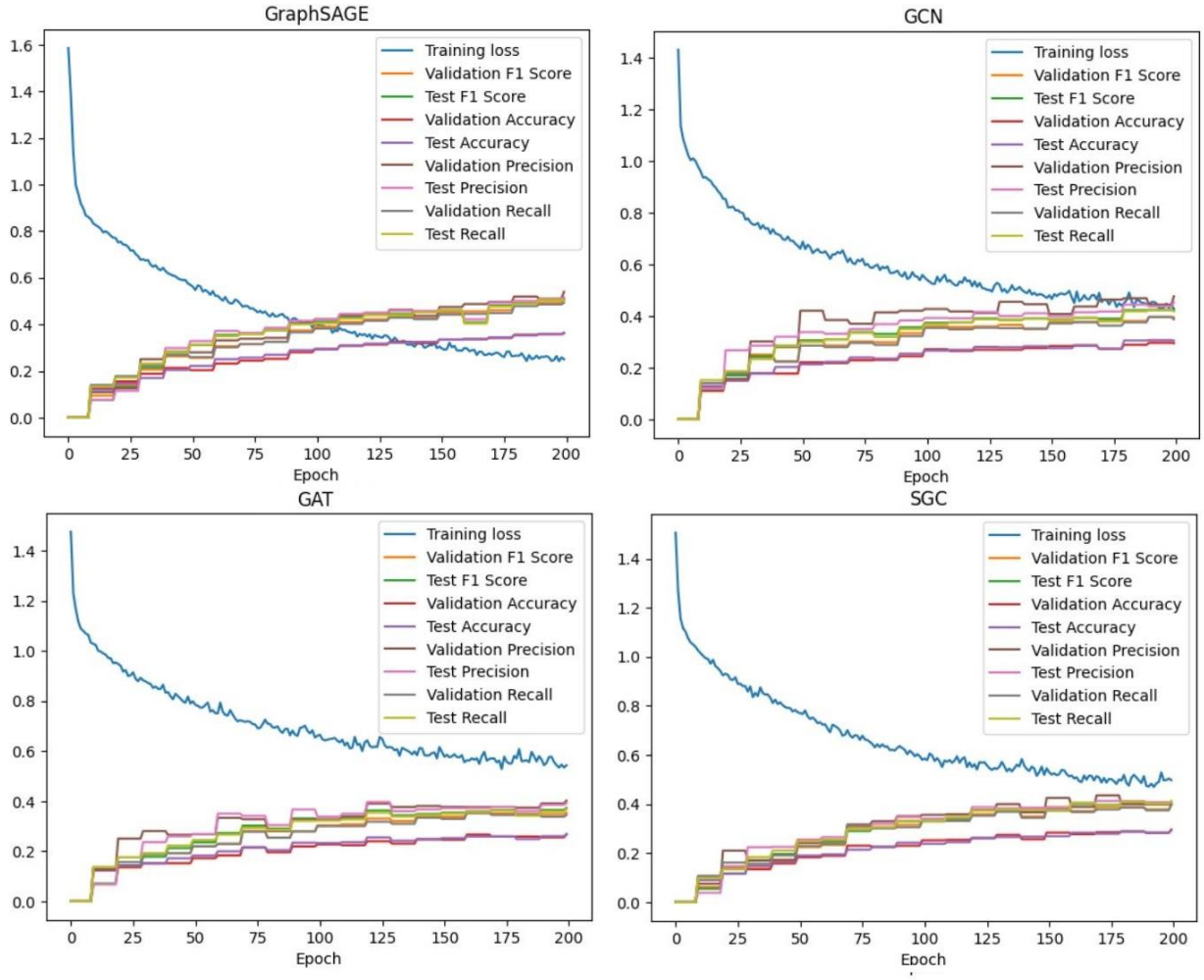

Fig 4. Training Loss and Model Performance on Validation and Test Sets over Epochs while using Industry Type as a Node Feature

V. Final Thoughts

The outcomes that we got in the task of binary link prediction reveal a satisfactory level of accuracy, highlighting the effectiveness of the models in deciding on the presence or absence of links. However, it is very important to recognize the detailed nature of the multiclass link prediction task, where the challenge increases due to the introduction of various link types. The results in this context may not match the performance achieved in the binary link prediction. The complexity occurs from the unequal distribution of relationships across different types, leading to a significant class imbalance that has affected the model's ability to generalize effectively. Furthermore, the decision to use a subset

of 2000 samples, because of the resource constraints, adds another layer of complexity. Increasing the dataset by including all available examples shows potential for improving the model's understanding of different relationships. It is worth noting that the observed trends, including a downward trajectory in training loss and concurrent upward trends in training and validation metrics (accuracy, recall, precision, and F1 score), indicate the potential for further model improvement with extended training epochs. This indicates that with more thorough training, the models could better grasp the complex relationships within the data, potentially resulting in improved predictive performance.

VI. Conclusion

In conclusion, our exploration into link prediction tasks within a graph-based framework has revealed insightful findings and considerations. Using the industry type as a node feature helps reduce bias that can occur when instead we use the node degree as a node feature. This approach introduces meaningful information that augments the model's understanding of node importance, leading to improved performance.

The choice of using the link type as an edge label, although resulting in decreased model scores, offers a more authentic representation of the dataset. The detailed relationships captured by considering link types provide a richer context, even at the expense of some predictive metrics. This compromise shows the significance of making model decisions that consider all the complexities present in real-world data.

It is crucial to acknowledge the limitations that are caused by the size of our training dataset. With a sample size of 2000, the models demonstrate good performance, however increasing the dataset could increase the model performance. A larger dataset not only aids in model generalization but also gives us a better understanding of the relationships present in the data. Our observations regarding the trajectory of training metrics show the potential benefits of extending the training duration. The downward trend in training loss and the upward trajectory of accuracy, recall, precision, and F1 score across epochs indicate that further training iterations could lead to enhanced model performance. This trend suggests that the models are continually learning and adapting to the complexities of the given task.

References

[1] Xiaowei Huang, Kaida Ning *Graph Neural Networks: Architectures, Advances, and Applications*, 2021

[2] Thomas Kipf *Graph Convolutional Networks*, 2017

[3] William L. Hamilton, Rex Ying, and Jure Leskovec *Graph Representation Learning*, 2021

[4] Amr Ahmed, Eric Xing *Graph Representation Learning*, 2018