MACHINE LEARNING METHODS FOR SOCIAL NETWORK ANALYSIS

Justin Baltazar
University of Massachusetts Amherst
Amherst, MA
jmbaltazar@umass.edu

Lucy Bodtman
University of Massachusetts Amherst
Amherst, MA
lbodtman@umass.edu

Manan Talwar
University of Massachusetts Amherst
Amherst, MA
mtalwar@umass.edu

ABSTRACT

The use of Social Network Analysis to examine the relationships within communities of people can be an invaluable tool to social scientists for understanding social processes and dynamics. In the recent times, Social Network Analysis methods have been applied in a variety of fields such as ecology, public health, life and social sciences. In this paper, we present an analysis of state-of-the-art machine learning methods that have been developed recently to identify their applications as well as potential in large scale Social Network Analysis. Our experiments show that we can use the low dimensional vectors and features generated by Machine Learning-based Social Network graphs to predict behaviors of individuals as well as communities. Based on our results, we propose that several forms of machine learning can find practical applications in various domains of social and life sciences.

1 Introduction

Social Network Analysis (SNA) refers to the process of investigating social structures through the use of networks and graph theory. With the rise of massive collection of data through the internet, social sites and large scale databases, the recent times have seen a rise in research aimed at employing Social Network Analysis to analyze characteristics of large scale social networks and understand social processes and their effects on the society [7]. Social Network Analysis offers deep insights into activities and the behaviours of entities which can be used for predictive modeling of complex real world social phenomena such as disease transmission and spread, misinformation and hate speech propagation through social networks, user classification and clustering and sentiment analysis.

There is a wide variety of tools available for Social Network Analysis. Traditional approaches to Social Network Analysis have been based on graph theoretic properties of networks. Recent developments in the discipline have focused on the application of machine learning to understand social networks [1]. The underlying idea is that Machine Learning enables machines to recognize patterns and draw conclusions with minimum human interference iteratively. Preliminary research has indicated that Machine Learning methods for Social Network Analysis have several advantages over traditional methods [1] [2]. One major class of Machine Learning models that has gained widespread attention for its applications in Social Network Analysis is Network/Node Embedding models [3].

Node Embedding models are a relatively recent phenomena [2]. These are powerful machine learning models that are designed specifically for graph-based machine learning. At its core, the process of node embedding is a transformation of nodes of a graph into a set of vectors while preserving important properties of the graph. In practice, this is achieved by defining similarity metrics such as dot products, euclidean norm, and the Frobenious norm. Model hyper-parameters are then tuned to ensure that similarity in the embedding space approximates similarity in the original network. This means that an accurate node embedding can capture the graph topology, node-to-node relationship, and other relevant information about the graph, its subgraphs, and nodes [3].

Motivated by the recent developments highlighting the vast potential of Machine Learning as a tool for Social Network Analysis, we propose the following research questions:

1. *RQ-1: What are some Node Embedding models that can be used for Social Network Analysis?*
2. *RQ-2: What are the underlying mathematical principles behind these Node Embedding models?*
3. *RQ-3: How do these Node Embedding models present advantages over traditional methods of Social Network Analysis?*

2 Background

Prior work in the domain [6], [7] has established that classical approaches such as regression and classification are not sufficient for modeling social network characteristics. There are several reasons for this. Most Machine Learning models cannot deal with graph based inputs. Further, the graphs analyzed in large scale Social Network Analysis can get arbitrarily large, presenting challenges to store and compute on them efficiently. Each node in such a graph can have a very large number of features and properties. As a consequence, Machine Learning models need to deal with high dimensional space and geometry.

2.1 Network Embedding Models

Network Embedding models are a classic solution to the problems posed by the high dimensionality of social network data. By embedding, we mean mapping each node in a network into a low-dimensional space which gives us insights into nodes' similarity and network structure. The node embedding process consists of three steps [3]:

1. Define an encoder i.e. a mapping from nodes to low dimensional vector embeddings.
2. Define a node similarity function i.e. a measure of similarity in the original network which specifies how the relationships in vector space map to the relationships in the original network.
3. Optimize the parameters of the encoder so that similarity of nodes in the network approximate the similarity between node embeddings in the low dimensional space.

2.2 Commonly used Network Embedding Models

Despite being recent, Network Embedding models are being used in several applications. In this section, we present an overview of some commonly used Network Embedding methods.

2.2.1 DeepWalk

DeepWalk [5] is a graph neural network method that uses a randomized path traversing technique to provide insights into localized structures within networks. It does so by utilizing these random paths as sequences that are then used to train a the classic Skip-Gram Language Model commonly used in Natural Language Processing applications for word prediction. These random paths are generated in an extremely simple manner. Starting from the target root, the algorithm randomly select a neighbor of the node, and adds it to the path. This process is continued iteratively until the desired number of steps has been taken. The Skip-Gram algorithm then tries to predict the neighbors around the target node within our network.

2.2.2 Walklets

Walklets [6] is a procedure that uses random walks to approximate the pointwise mutual information matrix obtained by individual normalized adjacency matrix powers. These are all decomposed by factoring powers of the adjacency matrix A and the embeddings are concatenated together. Similar to DeepWalk, the network is modeled through a series of truncated random walks started at each node. Additionally, we choose to skip some of the nodes in the random walk. In this way, we form a set of relationships which are sampled from successively higher powers of A. Finally, the Skip-Gram loss function is used for optimization.

2.2.3 Geometric Laplacian Eigenmap Embedding

GLEE [9] is a linear algebraic method that arises from spectral graph theory. The procedure extracts the eigenvectors corresponding to the largest eigenvalues of the graph Laplacian. These vectors are used as the node embedding. The procedure is as follows: Given a graph G, consider its Laplacian matrix L. Using singular value decomposition we may write $L = SS^T$ for a unique matrix S. Define S^d as the matrix of the first d columns of S. If i is a node of G, define its d-dimensional Geometric Laplacian Eigenmap Embedding (GLEE) as the ith row of S^d.

3 Methods

For our first and second research questions, we conducted an extensive literature review to identify Network Embedding models that can be especially useful in Social Network Analysis. For our third research questions, we developed implementations in Python for each of the network embedding methods described in the previous section.

3.1 Literature Review

We searched for relevant academic literature on Google Scholar, IEEE and Frontiers in Big Data. We defined the inclusion criteria to accept only empirical studies that reported validated statistically significant conclusions.

During the selection process, we first performed keyword search using 'Social Network Analysis', 'Machine Learning' and 'Network Embeddings'. Then, we manually reviewed the abstracts of the results for consideration. We selected 19 journal articles that we read and assessed for inclusion. Based on the inclusion criteria, we included 6 in our review [5], [9], [6], [7], [2], [1].

3.2 Code Implementations and Analysis

We used the Karateclub API [8] for running the embedding models. We split the data obtained from node embedding and used 20% for testing and 80% for training. We then used Logistic Regression and Random Forest classifiers from scikit-learn to analyze three datasets available open source via Stanford SNAP [4]:

- GitHub dataset $(V = 37700, E = 289003)$: A social network where nodes correspond to developers who have starred at least 10 repositories and edges to mutual follower relationships. The task is to classify nodes as web or machine learning developers.
- Deezer dataset $(V = 28281, E = 92752)$: A user-user network of European members of the music streaming service Deezer. The links represent mutual friendships of the users. The task is the classification of the users' gender.
- Twitch dataset $(V = 7126, E = 35324)$: User-user networks where nodes correspond to Twitch users and links to mutual friendships. The task is classification of whether a streamer uses explicit language.

4 Results

4.1 Literature Review

Analysis from literature review identified that nearly any Node Embedding can be successfully used to get low dimensional representations of graphs that can be used for Social Network Analysis. Current state-of-the-art methods that have been employed for Social Network Analysis and analyzed in academic settings are as follows:

- DeepWalk
- Walklets
- Geometric Laplacian Eigenmap Embedding
- GraRep

It is worth mentioning that while GraRep is backed by the strongest evidence, it is impractical to use on large scale social networks and hence, it does not effectively present any practical advantage in Social Network Analysis. We summarized the underlying mathematical principles for each of the remaining Network Embedding models in this list in 2.2.

4.2 Code Implementations

KarateClub provides hyper-parameter tuned settings that can be imported into Network Embedding models to evaluate their performance. We used the default settings employed by the authors who contributed their models to KarateClub to run implementations of DeepWalk, Walklets and GLEE on the three datasets as described in 3.2.

4.2.1 Twitch Dataset

We report the AUC observed across six different combinations of Network Embedding models followed by supervised classification in Table 1. We also report the confusion matrix and a classification probability plot for the best performing combination in Figure 1.

Table 1: AUC Report for Twitch Dataset

Graph Embedding Method	Logistic Regression AUC	Random Forest AUC
GLEE	0.5004	0.5383
Walklets	0.5684	0.5653
Deepwalk	0.5718	0.5597

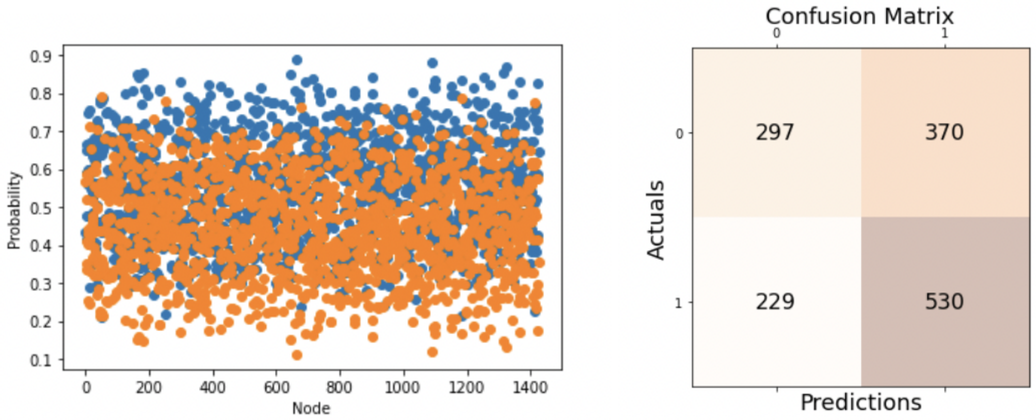

Figure 1: Classification Visualization and Confusion Matrix for Twitch Dataset

4.2.2 Github Dataset

We report the AUC observed across six different combinations of Network Embedding models followed by supervised classification in Table 2. We also report the confusion matrix and a classification probability plot for the best performing combination in Figure 2.

Table 2: AUC Report for Github Dataset

Graph Embedding Method	Logistic Regression AUC	Random Forest AUC
GLEE	0.5000	0.7064
Walklets	0.7821	0.7688
Deepwalk	0.7604	0.6112

4.2.3 Deezer Dataset

We report the AUC observed across six different combinations of Network Embedding models followed by supervised classification in Table 3. We also report the confusion matrix and a classification probability plot for the best performing combination in Figure 3.

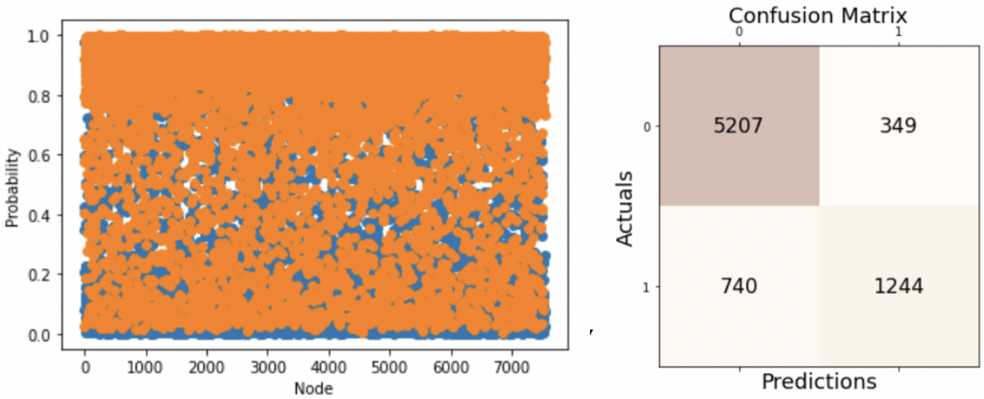

Figure 2: Classification Visualization and Confusion Matrix for Github Dataset

Table 3: AUC Report for Deezer Dataset

Graph Embedding Method	Logistic Regression AUC	Random Forest AUC
GLEE	0.5015	0.5357
Walklets	0.5280	0.5418
Deepwalk	0.5263	0.5095

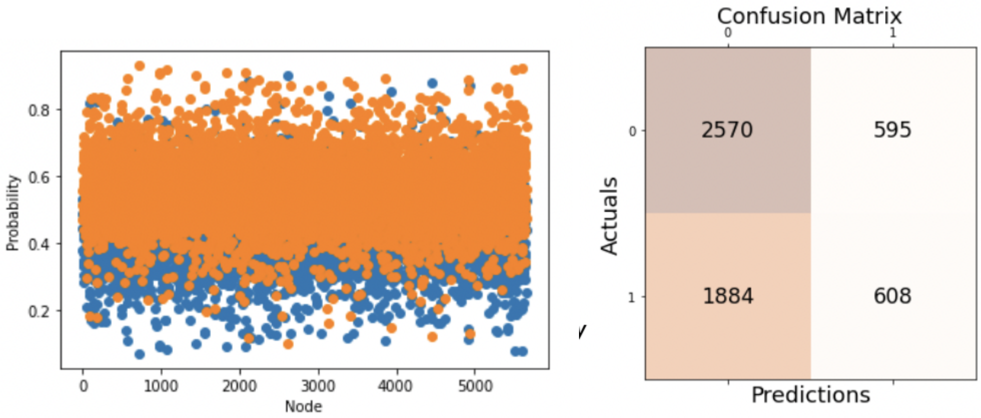

Figure 3: Classification Visualization and Confusion Matrix for Deezer Dataset

4.3 Inferences

Based on our experiments, we report the following inferences:

- Random walk based methods (DeepWalk and Walklets) consistently outperform linear algebraic methods (GLEE) in our study as seen in 4.2.
- In our experiments, Walklets emerges to be the best performing model over 2 of the 3 datasets in the study (Github and Deezer).
- Comparing our results with prior literature in the discipline, we observe that AUC over testing set in our experiments is in tandem with observations in the existing literature in the discipline.
- We report that in our experiments, GLEE tends to work well with ensemble Random Forests but works poorly in Logistic Regression.

5 Conclusion

In this study, we explored Network Embedding models as a tool for Graph-based Machine Learning that can be used for large scale Social Network Analysis. We identified three major Network Embedding models that have been used for Social Network Analysis in the recent time and evaluated their performance through our experiments. We observed that none of the models are universally ideal. There are theoretical limitations to the performance and convergence of each of these methods. We noted that it is crucial to understand that Network Embeddings are approximations to the data. Network Embeddings can and do lead to information loss and hence, not all graph characteristics are perfectly preserved by Network Embedding procedures. We observed that the characteristics as well as the struture of the data affects the performance of the Machine Learning models and hence, it is imperative to select a Network Embedding model suited to the specific type of data one desires to analyze. We noted that our promising findings are in tandem with those reported by other researchers in this discipline. Our results identified research potential in the field of Social Network Analysis via Network Embeddings. We invite future researchers to explore the vast potential of Network Embeddings in Graph-based Machine Learning.

Author's Contribution

All authors contributed to the literature review. Justin Baltazar prepared the Abstract. Lucy Bodtman prepared the Introduction. Manan Talwar prepared the Background and Methods. Justin Baltazar conducted analysis on the Github dataset. Lucy Bodtman conducted analysis on the Deezer dataset. Manan Talwar conducted analysis on the Twitch dataset. All authors contributed to the Results section. Lucy Bodtman prepared the Conclusions section.

References

[1] BALAJI, T., ANNAVARAPU, C., AND BABLANI, A. Machine learning algorithms for social media analysis: A survey.

[2] HAYAT, M. K., DAUD, A., ALSHDADI, A. A., BANJAR, A., ABBASI, R. A., BAO, Y., AND DAWOOD, H. Towards deep learning prospects: Insights for social media analytics. *IEEE Access 7* (2019), 36958–36979.

[3] LESKOVEC, J. Node representation learning.

[4] LESKOVEC, J., AND KREVL, A. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[5] PEROZZI, B., AL-RFOU, R., AND SKIENA, S. Deepwalk: Online learning of social representations, Jun 2014.

[6] PEROZZI, B., KULKARNI, V., CHEN, H., AND SKIENA, S. Don't walk, skip! online learning of multi-scale network embeddings, Jun 2017.

[7] ROZEMBERCZKI, B., ALLEN, C., AND SARKAR, R. Multi-scale attributed node embedding, Mar 2021.

[8] ROZEMBERCZKI, B., KISS, O., AND SARKAR, R. Karate Club: An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)* (2020), ACM, p. 3125–3132.

[9] TORRES, L., CHAN, K. S., AND ELIASSI-RAD, T. Glee: Geometric laplacian eigenmap embedding, Mar 2020.