LINK PREDICTION IN SOCIAL NETWORKS

A dissertation

submitted in fulfillment of the requirements for the
award of the degree of

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

HIMANSHU TIWARI

(2K19/CSE/09)

Under the Supervision of

MR. SANJAY KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi, India-110042

AUGUST, 2021

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi, India-110042

CANDIDATE'S DECLARATION

I, **HIMANSHU TIWARI**, Roll No. **2K19/CSE/09** student of MTech Computer Science and Engineering, hereby declare that the project dissertation titled **"Link Prediction in Social Networks"** which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of **Master of Technology**, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi
Date: 01-08-2021

Himanshu Tiwari
2K19/CSE/09

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi, India-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Link Prediction in Social Networks"** which is submitted by **HIMANSHU TIWARI**, Roll No. **2K19/CSE/09** Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of **Master of Technology**, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 12/08/2021

MR. SANJAY KUMAR
SUPERVISOR

ACKNOWLEDGEMENT

Mr. Sanjay Kumar, Assistant Professor, Department of Computer Science, Delhi Technological University, has my undying admiration and gratitude for his wisdom, vision, expertise, guidance, enthusiastic participation, and unwavering encouragement throughout the planning and development of this research project. I also appreciate his careful work in reviewing and polishing the manuscripts, without which this work could not have been completed.

I would also like to express my gratitude to members of the CSE Department's personnel for their prompt assistance and cooperation during the investigation. Finally, I owe and thank the Almighty for assisting me in this quest.

Place: Delhi **Himanshu Tiwari**
Date: 01/08/2021 **2K19/CSE/09**

ABSTRACT

In today's society, social networks play a significant role, with applications ranging from creating a more connected world to finding critical relationships in biological systems. The significant growth in the use of social networks has increased the need of recognizing node-to-node relationships even before they are formed. Several approaches for the task of link prediction utilizing various indices have been developed in the past. There has been a lot of work put into combining multiple indices utilizing machine learning techniques and analogies to the Law of Gravitation, with similarity measures serving as proxies for distance and popularity measures serving as proxies for mass. Merging different indices can improve overall link prediction efficacy, although only a few techniques have been proposed in the past. After integrating three popularity and four similarity metrics, we suggest the usage of a "***Histogram based Gradient Boosting Regression Tree***" for the task of link prediction in this work. ***Nature Inspired Approach using CC-CD***, has also been proposed which makes use of node embeddings and closeness centrality. Node Embeddings is a way of representing the high dimensional vector representation of graphs to a low dimensional vector. We have used the cosine distances of node embeddings as a proxy of distances and Closeness Centrality as a proxy of masses in Newton's Gravitational Law for prediction of new links.

KEYWORDS

Link prediction, node embeddings, social networks, complex networks.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

ML = Machine Learning

PCA = Principal Component Analysis

ICA = Independent Component Analysis

EDA = Exploratory Data Analysis

MFI = Matrix Forest Index

HGBLP = Histogram-based Gradient Boosting Regression Tree for Link Prediction

1-D = One Dimensional

2-D = Two Dimensional

SDNE = Structured Deep Network Embeddings

PDF = Probability Density Function

CDF = Cumulative Distribution Function

SVM = Support Vector Machine

KNN = K- Nearest Neighbors

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

TPR = True Positive Rate

FPR = False Positive Rate

TNR = True Negative Rate

FNR = False Negative Rate

MR = Miss Rate

Acc = Accuracy

Sen = Sensitivity

Spec = Specificity

AUC= Area under the Curve

ROC = Receiver Operating Characteristic

μ = Mean

σ = Variance

% = Percentage

CHAPTER 1 INTRODUCTION

There has been a recent surge in the complexity of social networks related owing to the success of social media, collaboration, and epidemiology. The research studies of social networks have always been an area of interest. Social network analysis deals with mining and analyzing the voluminous data and predicting valuable patterns, which can be useful for numerous real-life applications like community detection [26], influence maximization [24, 2], information diffusion [26] and fake news detection. Link prediction is one of the popular research topics in the field of social network analysis which deals with predicting links in social networks even before formation. The social networks are generally represented by a set of nodes and edges, where nodes are people and edges represent relationships. These social networks can be a set of scientists and their relationships representing co-authorship, a set of employees related by their collaboration on projects, or relatively new but interesting examples of Facebook friends [1]. Most of the tasks related to social networks are algorithmically complex owing to the large size of the underlying network structures. The analysis of the results also becomes complex because the defined algorithms generally use heuristics. Needless to say, social networks are highly dynamic. New relationships and new nodes are created at a very fast speed in a typical social network. Monthly active users increase by ten percent year over year and Facebook is accessed by over 56% of the world's active internet users. Social networks are important for tasks such as transmitting information, predicting the outcomes of an event, and predicting new relationships. Most of the tasks related to social networks are algorithmically complex owing to the large size of the structures. The analysis of the results also becomes complex because the defined algorithms generally use heuristics. The algorithms can be analyzed statistically on previous data but the predictions on new data may not be a hundred percent sure.

Link prediction involves the prediction of relationships even before they are formed. This task is of great importance because we can predict various aspects of the networks with great accuracy like the suggestion of friends on a social network, predicting the relationships contributions of scientific authors, and even underground relationships between terrorists. We can even predict or find out relationships between the various actors of some act using link prediction techniques. These techniques can also be used for personalized recommendations for certain products on e-commerce websites and personalized advertisements on the web.

There have been several algorithms produced from time to time for link prediction problem like the similarity-based link prediction algorithms, maximum likelihood-based link prediction algorithms, and probability-based link prediction algorithms. Similarity methods are based on the hypothesis that the greater similarity (having several common features) measure of nodes results in better chances of their future link formation [20]. Maximum likelihood methods are based on some assumptions and principles the networks are based upon. Based on these assumptions and principles, detailed rules are devised and the likelihood of presently non-existing link to form in the future is calculated. These assumptions involve that most networks are organized in hierarchies on different levels [21]. These methods are essentially based on the structural characteristics of the networks like hierarchies or communities [22]. Probability-based methods try to find a function that abstracts the underlying network and fits the available data to create an observed network. Once this model is created, the future links can be predicted using conditional probability and the parameters the probabilistic function is based upon [23, 25].

In the recent past, there have also been presented some algorithms inspired by the laws of physics. These approaches primarily make use of some physical formula and use the analogy of different measures in the context of social networks. The most commonly used formula is the Law of Gravitation formula. This formula states that the gravitational force between two masses is proportional to the product of their respective masses and inversely proportional to the square of the distance between them. The researchers have used the analogy of treating the bodies under consideration as nodes of the network and the force due to gravitation as the measure of the link prediction task.

Single similarity indexes are unable to provide such exact findings, therefore, few researchers have also introduced the idea of a merger, however, there are only a few approaches that use this technique. The merging of several indices can increase the overall effectiveness of link prediction, according to existing merger-related research. This is because it can get through the problem of a single index only being suitable for networks with identical structural properties. The existing research has also established that the merging of indexes is not the primary reason for improved outcomes; rather, it is also the selection of all those indices that affects link prediction results [27, 28].

In this work, we suggested a method called ***Histogram-based Gradient Boosting Regression Tree for Link Prediction (HGBLP)*** based on the fusion of three popularity and four similarity measures based on the aforesaid study. The method appears to yield superior results in theory, as various indices represent different network features. We also provide an approach called ***Nature Inspired Approach using CC-CD*** for the link prediction task using the nature inspired approach along with node embeddings. Node Embeddings capture the essential features of the network while abstracting the underlying network. These embeddings represent the nodes in a low dimensional vector, from which we can find the dissimilarity of nodes using the cosine distances. We propose using these measures along with the closeness centrality to plug in the gravitation law and attempt to formalize the measures in a much more relatable way.

CHAPTER 2 LITERATURE REVIEW

The link prediction task involves the prediction of new relationships that may be created over time. More formally, suppose we have a social network's snapshot at current time t, we are interested in accurately predicting the edges that are not currently present at time t but are present at time t+h.

There have been many link prediction approaches defined in the past. The approaches can be classified into the following categories [3]:

Similarity-Based link prediction algorithms

These methods take two nodes representing a non-existent link and assign a score to the pair. These similarity scores can be as simple as the number of common neighbors between them. This approach has been analyzed from time to time in various studies and presented a positive correlation between the link formation and higher value of common neighbors [18, 19]. Link prediction algorithms based on similarity assign a score S_{xy} for each pair of nodes x and y. This score is defined as similarity/proximity between x and y. All the pairs which do not have a link between them at time t are ranked based on their score S_{xy}. Similarity indices can be very simple or very complex. Some similarity indices may be well suited for some networks while they may completely fail for some other social network. These similarity indices are based on the structural similarity of the nodes. These similarity measures can be derived by only considering the local structure or global structure.

Maximum likelihood Based link prediction algorithms

The maximum likelihood algorithms are recently introduced for link prediction. These algorithms tend to obtain detailed rules and parameters by observing the structures and estimating the maximum likelihood. New links are predicted using the rules and parameters. These algorithms are generally very slow for even a few thousand nodes.

Probability-based link prediction algorithms

There have been various probability-based models proposed in the past. These include Probability Relational Models (PRMs) such as the Relational Bayesian Network, Probabilistic Entity-Relationship models (PERMs), and Stochastic Relational Models [23, 25].

Nature inspired link prediction algorithms

In the recent past, various authors have also proposed several nature inspired approaches for link prediction. These approaches primarily use laws of physics and try to plug in various network measures in the formula. Bastami, Mahabad and Taghizadeh proposed a gravitation law-based link production model [10]. They have used community and network information for the reduction of the prediction space. Wahid-Ul-Ashraf, Budka, and Musial used many combinations for similarity and centrality of nodes to replace the mass and distances in the gravitation law [11]. They have used and evaluated a combination of three node centrality measures as mass and thirteen dissimilarity measures as distance in Newton's Gravitation Law. This results in the evaluation of over 50 combinations in their work. The results in their work suggest that the predictor using the combination of Closeness Centrality and Matrix Forest Index yielded the best overall results in AUC. In this work, we have also tried to benchmark our results against this combination and Matrix Forest Index. Kumar, Chaudhary, Kedia, and Singhal proposed various new measures and plug them in the gravitational law to further outperform these approaches [12].

Node Embeddings

Network Embedding, the problem of learning low dimensional vectors from traditional representation of networks has been studied extensively in the previous years. Interestingly, network embedding also captures the topology information contained in the networks. Notable developments in this regard are DeepWalk and node2vec [5]. These strategies use natural language processing algorithms like Word2Vector [4]. Higher-Order Proximity Preserved Embedding (HOPE) [6] approach generates embedded matrices based on matrix properties like the Katz similarity [7] measure. Structured Deep Network Embedding (SDNE) [8] was the first technique to be able to effectively capture the highly non-linear network structure in network embedding techniques. It is a semi-supervised deep learning-based method that used Laplacian Eigenmaps [9] and can preserve local as well as global structures of the network. These network embeddings can be used for link prediction tasks in social networks as shown by the respective authors.

CHAPTER 3 NODE EMBEDDINGS

Many real world applications can be modelled as a graph problem where entities are connected to each other through links. Analyzing these structures yields interesting results and properties of the underlying entities. These networks become very complex consequently tasks such as classification, clustering and prediction require high computational complexity and lack of parallelizability becomes a bottleneck when traditional representations of the network are used. Machine Learning algorithms for the above-mentioned tasks have seen tremendous growth in previous years and have been very powerful in certain applications. Most of these machine learning algorithms cannot be applied to traditionally represented networks. The above problems predominantly arise due to the fact that the nodes in traditional network forms are interdependent and the representation of a node would require a vector of high dimension. Network Embedding, the problem of learning low dimensional vectors from traditional representation of networks has been studied extensively in the previous years. Interestingly, network embedding also captures the topology information contained in the networks.

Node2Vec

Natural Language Processing involves similar tasks of learning word representations in order to capture the meaning of words, given the co-occurrence data in the form of text. Development of Word2Vector significantly improved the task of creating word embeddings, ie. fixed low dimensional vector representation of words. Since, nodes can be viewed as words and walks on the network can be viewed as short sentences, it formed the basis for random walk-based approaches for network embedding.

DeepWalk is an unsupervised way of learning latent representation of networks which is adaptive, scalable and online unlike most of the previous works. Uniform random walks of fixed length are used in DeepWalk which consequently provides no control over the neighborhood sampling. This disadvantage resulted in the development of more controlled strategies.

The sampling methodology of neighborhoods in the random walk guide the diversity of connectivity patterns in the final embedding vectors. Classic search strategies i.e breadth-first and depth-first sampling reflect extreme representations having homophily and structural equivalence respectively. Homophily hypothesis must generate close embeddings for the nodes that are highly interconnected, whereas structural equivalence must generate similar embeddings for nodes having similar structural roles (for example, a hub).

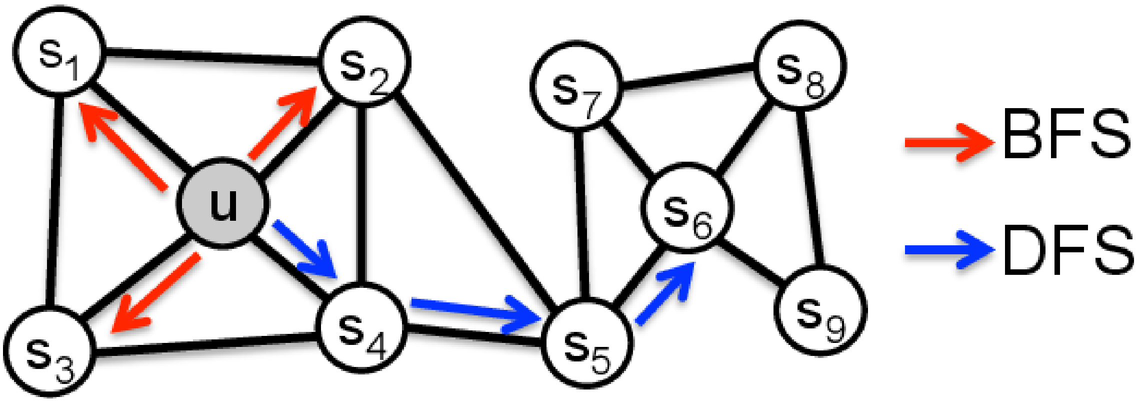

Fig. 1.DFS and BFS explore the graph differently.

The development of node2vec provided a flexible and biased neighborhood sampling strategy that allowed exploration in DFS as well as BFS fashion.

Node2vec is a semi-supervised algorithm that provides a way of controlling the neighborhood sampling via two parameters namely the return parameter and the in-out parameter. These parameters provide a search bias to the random walks for sampling the next node. A random walk currently at node v, which just traversed an edge (t, v) decides the next step based on the transition probabilities on all edges that lead from v.

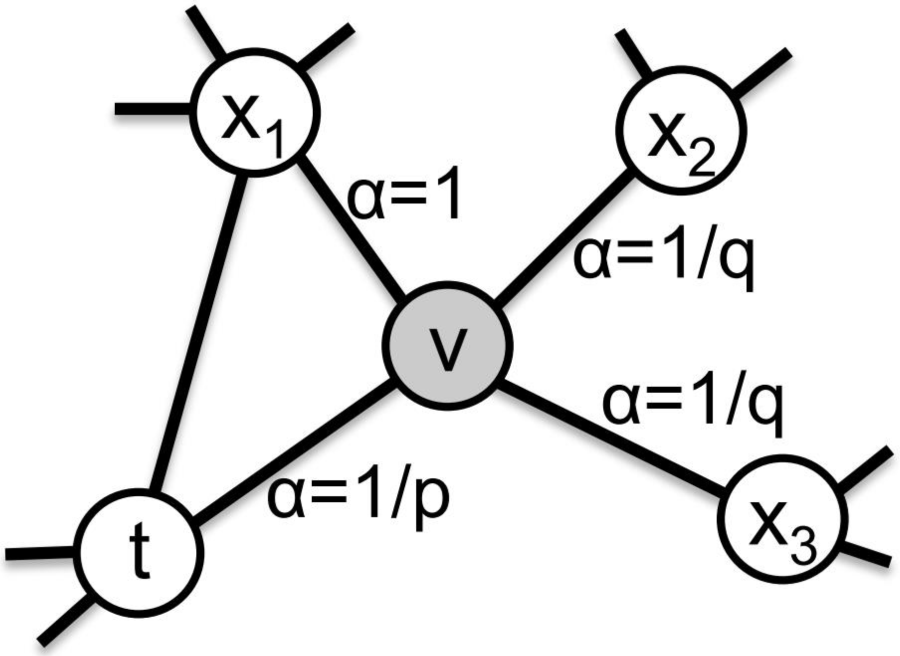

Fig. 2.Controlling the sampling of Node2Vec using different parameters.

The transition probability of an edge is defined as the product of the edge weight and search bias α. Value of α is based on the shortest distance between nodes t and x. Since the shortest distance must be one of {0, 1, 2}, search bias, α also has three possible values.

Shortest distance of 0 guides the walk back towards the node t in following two hops, thus it is controlled by the return parameter, p. A high value of p ensures that the probability of sampling an already visited node is lower in the next two steps. A low value of p is used to keep the walk “local” to the starting node.

Sampling of nodes with shortest distance of 2 is controlled by the in-out parameter, q. A high value of q ensures the sampling bias which is “local” and hence approximates the BFS behavior. On the other hand a lower q provides sampling which is biased towards the DFS behavior.

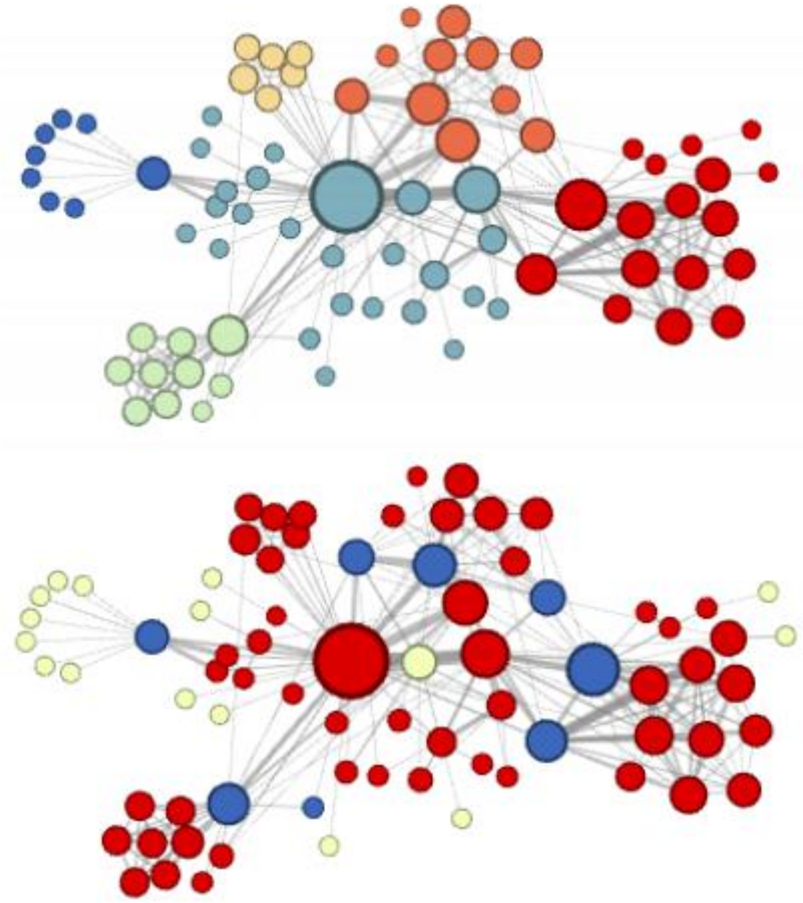

Fig. 3. Visualizations of Les Misérables network generated by node2vec reflecting homophily (top) and structural equivalence (bottom).

HOPE

For network embedding, factorization-based techniques can be used, in which we represent the network as a matrix and then factorize the matrix based on matrix properties to obtain the embedded matrix. HOPE is one of these methods, in which the Katz similarity matrix is used to create a low-dimensional vector representation of our network. HOPE's fundamental goal is to maintain asymmetric transitivity in directed graphs. No approach could manage asymmetric transitivity in directed graphs until HOPE was suggested. In undirected graphs, transitivity is symmetric, whereas in directed graphs, it is asymmetric.

On our input graph, we use Katz proximities, which is a high-order proximity measure in graphs that can reflect asymmetric transitivity. If there is a higher possibility of having a path between vi and vj, the value of Katz proximity will be higher for that edge.

Structured Deep Network Embedding (SDNE)

SDNE was the first network embedding approach to be able to efficiently capture highly non-linear network structure. It is a semi-supervised deep learning-based strategy that can preserve both local and global network topologies. It's a semi-supervised method since it contains two parts: supervised and unsupervised, that work together. It employs autoencoders, which are unsupervised neural networks that are utilized to represent the same input data in a considerably more dense fashion. Through an ongoing optimization process known as training, neural networks can steadily approximate a function that maps inputs to outputs. In a neural network, there may be a hidden layer with fewer nodes than the input layer, but it can be utilized to represent the same information as the input data. We name this network an autoencoder because it encodes a more dense representation of input data. In autoencoders, we configure the output to be the same as the input. Later on, the autoencoder outputs the same input as the output.

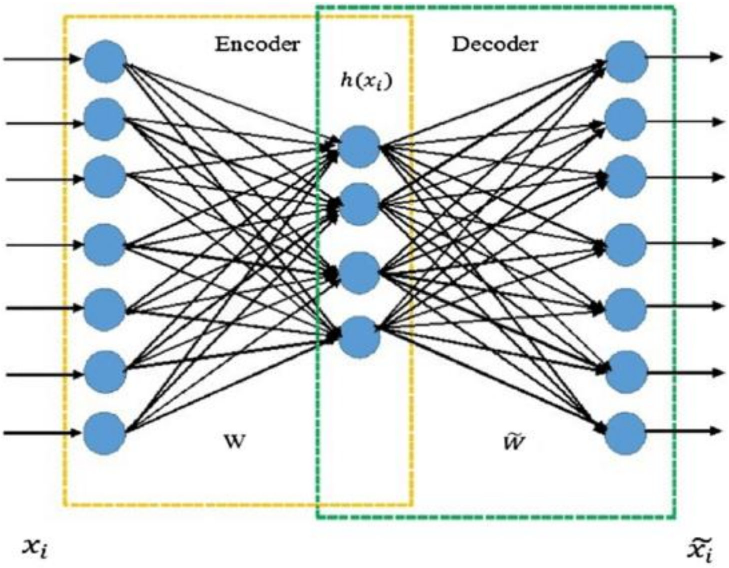

Fig. 4. Structured Deep Network Embedding (SDNE)

SDNE uses autoencoders to get the non linear network's embedded structure, and the same autoencoder can be used to reconstruct the neighborhood. That is, nodes in close proximity will have similar latent representations. Second-order closeness is thus kept. The presence of links between nodes in the rebuilt network indicates their similarity, while the absence of links does not always imply dissimilarity. As a result, the reconstructed network does not yet replicate the original source network accurately. To solve this problem, we employ Laplacian Eigenmaps, which have a cost when similar nodes in the embedding space are mapped far apart. To preserve first order proximity, Eigenmaps employs supervised information from the adjacency matrix. Furthermore, due to the network's sparsity, the number of zero elements in the adjacency matrix is significantly greater than the number of non-zero elements. We then apply a greater penalty for non-zero element reconstruction errors than for zero element reconstruction errors.

CHAPTER 4 NATURE INSPIRED APPROACH USING CC-CD

One important characteristic of Node Embeddings is that we can find the measure of similarity/dissimilarity using the cosine distances of the low dimension vectors. This distance when used with nature inspired approaches along with other centrality measures could be of special significance.

We would be using Node Embeddings to further the notion of distance in the classical law of gravitation formula, which is given by (1).

$$F \propto G * \frac{M_1.M_2}{D(i,j)^2} \tag{1}$$

Using (1) and taking the value of G as 1, we define a new quantity $F'(v_i, v_j)$ as a measure for link prediction task for node pair (v_i, v_j), as the square root of F. M_i refers to the centrality measure and $D(i,j)$ refers to the distance/dissimilarity between the two nodes (v_i, v_j).

$$F'(v_i, v_j) = \sqrt{\frac{M_{\mathrm{i}}.M_{\mathrm{j}}}{D(i,j)^2}} \tag{2}$$

Realigning (2) and using subscripts for nodes (v_i, v_j) we get (3)

$$F'(v_i, v_j) = \frac{\sqrt{M_{\mathrm{i}}.M_{\mathrm{j}}}}{D(i,j)} \tag{3}$$

The previous methods describe the denominator in (1) as dissimilarity. In the original gravitation law, the denominator is represented as the distance between the two masses. To further the notion of distance, we propose to find the measure F’ as the probability of future link formation.

We propose the distance between the nodes in (3) to be the distance between the vector representation of the node embeddings. Since node embeddings contain the essential information about the structure in a lower dimension, we can find any of the possible distances such as Euclidean, Hamiltonian, and cosine distances.

Although, the algorithm can be used by plugging various combinations of embedding algorithms, distance measures, and centrality measures, for this research, D in (3) represents the cosine distances between the node embeddings produced by node2vec.

Node Embeddings were calculated using node2vec and the square root of the product of closeness centralities was used in the numerator.

Methodology

The previous methods describe the denominator in (1) as dissimilarity. In the original gravitation law, the denominator is represented as the distance between the two masses. To further the notion of distance, we propose to find the measure F' as the probability of future link formation.

We propose the distance between the nodes in (3) to be the distance between the vector representation of the node embeddings. Since node embeddings contain the essential information about the structure in a lower dimension, we can find any of the possible distances such as Euclidean, Hamiltonian, and cosine distances.

Although, the algorithm can be used by plugging various combinations of embedding algorithms, distance measures, and centrality measures, for this research, D in (3) represents the cosine distances between the node embeddings produced by node2vec.

Node Embeddings were calculated using node2vec and the square root of the product of closeness centralities was used in the numerator.

Datasets Used

The below real-world datasets were used for the analysis of the proposed solution. These datasets are available freely and the description of these datasets is given below:

1) Jazz Musicians: The nodes are represented by jazz musicians and an edge denotes those two musicians who have played together. [13] ("SNAP Datasets")

2) Les Misérables: This data represents the characters in Les Misérables. [13] ("SNAP Datasets")

3) Football: This data represents the network of American Football games played between colleges of Division IA during the regular season Fall 2000. [14] ("American College Football Dataset")

4) Dolphin: A social network of bottlenose dolphins. The links represent frequent associations between dolphins. [15] ("Dolphins Dataset")

5) gr-qc: Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network is a description of scientific collaborations between authors' papers in the category. There is an undirected edge between authors if they co-authored some paper. If multiple authors were involved, it produces a complete subgraph of those authors. [16] ("GR-QC Dataset")

The data used was preprocessed to bring in a common format having integer nodes. Since the edges are unweighted, we did not need to take weights into the account. The characteristic features of the above datasets are summarized in Table I. The datasets considered present a variety of subjects and have number of nodes as low as 52 to as high as 5242.

TABLE I. DATASETS USED FOR EVALUATION OF THE PROPOSED SOLUTION

Dataset	**Number of vertices**	**Number of edges**	**Average Degree**
Jazz Musicians	198	2742	27.6970
Les Misérables	77	254	6.5974
Football	115	613	10.6609
Dolphin	62	158	5.0968
gr-qc (general relativity and quantum cosmology)	5242	14496	5.5307

Evaluation Metrics

To evaluate the effectiveness of our link prediction algorithm using node embeddings in the nature inspired approach, we have used the area under the receiver operating characteristic (ROC) curve (AUC). AUC is a well-known measure that is used as the probability of giving a randomly chosen non-existent link a lower score than a randomly chosen missing link. Suppose, we perform N independent comparisons of the calculated scores from our algorithm, and out of these N comparisons, N’ times missing links have a higher score. Also, N’’ times the score is less than or the same, then the AUC value is defined by (4).

$$AUC = \frac{N' + 0.5N''}{N} \tag{4}$$

Results

The calculated AUC values using the nature inspired approach with closeness centrality and cosine distances between the node embeddings produced by the node2vec algorithm for the datasets are summarized in Table II. The AUC values for the link prediction task by the node2vec algorithm alone are summarized in Table III.

The chart in Fig. 5 shows the comparison between the calculated values by our proposed solution and the node2vec algorithm. The vertical axis corresponds to the area under the receiver operating characteristic (ROC) curve (AUC) and the x-axis corresponds to the different data sets used for experimentation.

It can be concluded from Fig. 5, that our proposed algorithm performed better than the node2vec algorithm alone on every dataset used during the experimentation. Thus, the findings suggest that, it is better to use node2vec along with closeness centrality rather than using it alone.

TABLE II. CALCULATED AUC VALUES USING EQ. 3 WITH CLOSENESS CENTRALITY AND COSINE DISTANCES BETWEEN NODE2VEC EMBEDDINGS

Dataset	Calculated AUC Value
Jazz	0.91651456489
Les Misérables	0.72689075630
Football	0.79558823529
Dolphins	0.500000000
gr-qc	0.9821018585

TABLE III. CALCULATED AUC VALUES USING NODE2VEC EMBEDDINGS

Dataset	Calculated AUC Value
Jazz	0.86707703927
Les Misérables	0.65126050420
Football	0.76339285714
Dolphins	0.5000000000
gr-qc	0.9237911430

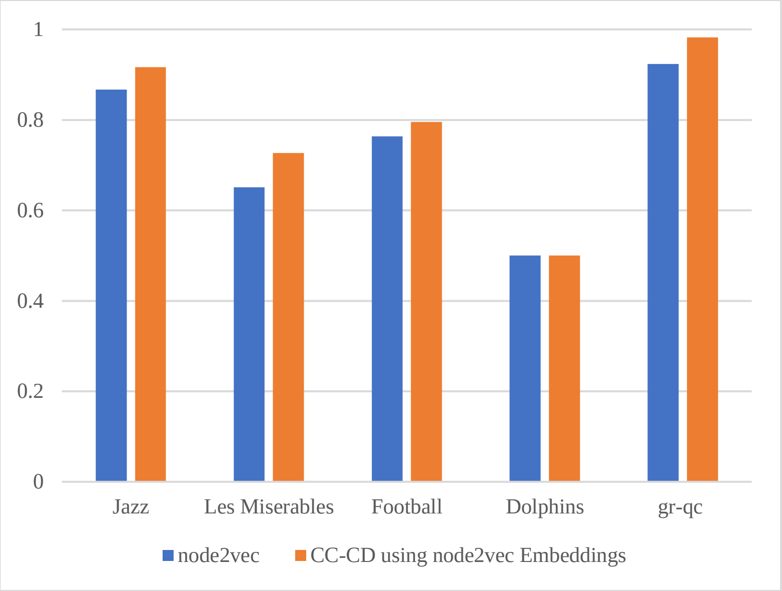

Fig. 5.Chart showing the comparison of AUC values using node2vec and using the nature inspired approach with closeness centrality and cosine distances using node2vec embeddings.

The AUC values for the link prediction task using the nature inspired approach proposed by Wahid-Ul-Ashraf, Budka and Musial. are summarized in Table IV. We have used their recommended pair of closeness centrality and the MFI in the gravitational law in (1). Equation (5) is used for the calculation. MFI [17] here means Matrix Forest Index similarity measure between two nodes (v_i, v_j).

$$F(v_i, v_j) = CC_i * CC_j * \ MFI^2 \qquad (5)$$

The chart in Fig. 6 shows the comparison between the calculated values by our proposed solution and nature inspired approach proposed by Wahid-Ul-Ashraf, Budka, and Musial. using Closeness Centrality and MFI (CCM). The vertical axis corresponds to the area under the receiver operating characteristic (ROC) curve (AUC) and the x-axis corresponds to the different data sets used for experimentation.

It can be concluded from Fig. 6, that our proposed algorithm performed better on every dataset used during the experimentation.

TABLE IV. CALCULATED AUC VALUES USING NATURE INSPIRED APPROACH – CCM

Dataset	Calculated AUC Value
Jazz	0.54000588
Les Misérables	0.668072188
Football	0.58263657
Dolphins	0.465494792
gr-qc	0.501860918

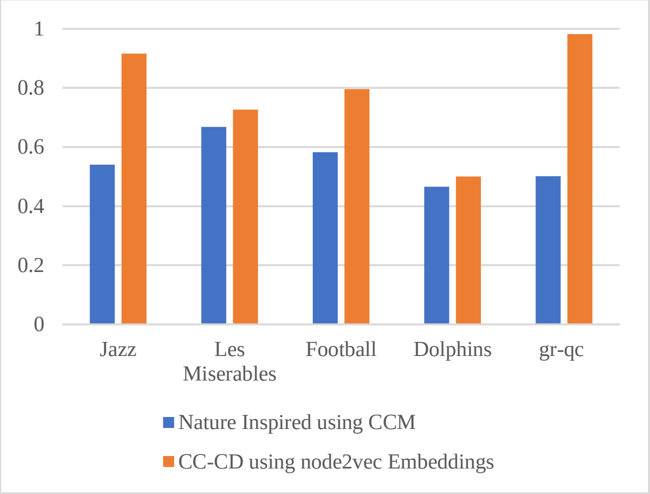

Fig. 6.Chart showing the comparison of AUC values using nature inspired approach with CCM and using the nature inspired approach with closeness centrality and cosine distances using node2vec embeddings.

The AUC values for the link prediction task using the Matrix Forest Index (MFI) are summarized in table V. MFI, a similarity measure between two nodes (v_i,v_j), is defined as the ratio of number of spanning forests such that v_i and v_j are contained in the same tree with v_i as the root to all spanning rooted forests of the main graph [17]. The spanning subgraph may not have all the edges of the graph but it has all the vertices of the graph. A forest is a graph having no cycles. A tree is a connected graph with no cycles.

We have compared our model with the MFI because it is recommended by Wahid-Ul-Ashraf, Budka, and Musial. with the nature inspired approach. In general, nature inspired approach with Closeness Centrality and MFI in Table IV seems to outperform the MFI in Table V, as can be seen in Fig. 8 comparing both the approaches.

We tested MFI results in Table V against our model's results in Table II and found that the use of cosine distances of node embeddings using node2vec along with the closeness centralities outperforms the Matrix Forest Index.

TABLE V. CALCULATED AUC VALUES USING MFI

Dataset	Calculated AUC Value
Jazz	0.516762005
Les Misérables	0.713526733
Football	0.585118327
Dolphins	0.461371528
gr-qc	0.50079692

The chart in Fig. 7 shows the comparison of our model with the Matrix Forest Index and it can be concluded from the results that the new model outperformed the MFI on all the datasets used for the experimentation.

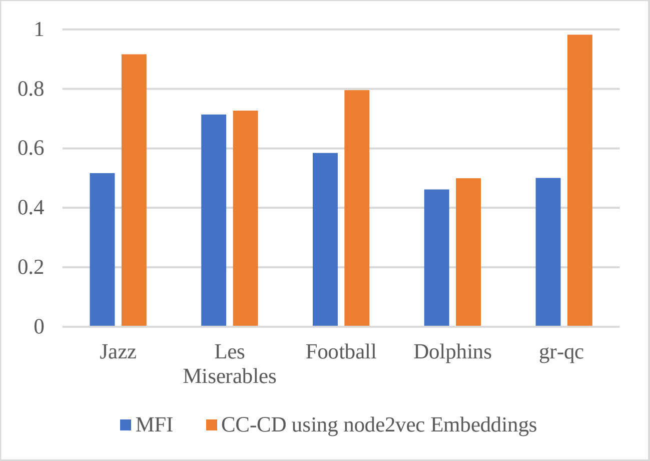

Fig. 7.Chart showing the comparison of AUC values using MFI and using the nature inspired approach with closeness centrality and cosine distances using node2vec embeddings.

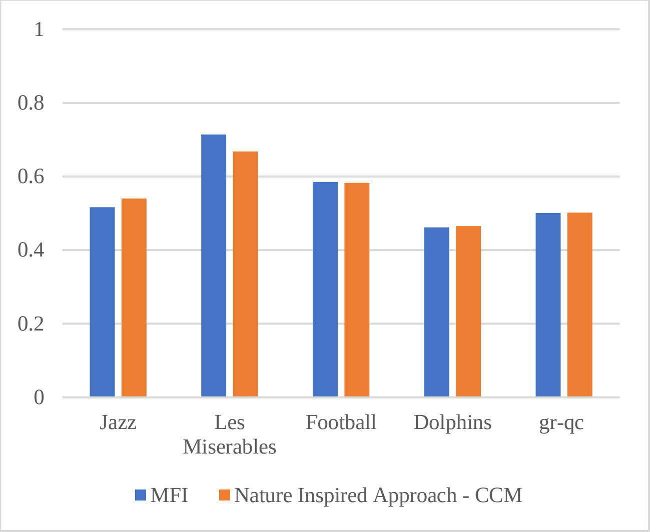

Fig. 8.Chart showing the comparison of AUC values using MFI and using the nature inspired approach-CCM.

Results

Based on the experiments conducted for the proposed work in this study, we can conclude that equation (3) outperforms the previous nature inspired approach as well as the individual algorithms used in the formulae (node2vec, MFI, and nature inspired - CCM), when used with cosine distances in node embeddings produced by node2vec instead of the proposed similarity/dissimilarity in the previous works. The future work for the study can explore the combination of different node embeddings in combination with different centrality measures.

CHAPTER 5 LINK PREDICTION IN SOCIAL NETWORKS USING HISTOGRAM BASED GRADIENT BOOSTING REGRESSION TREE

The link prediction challenge entails predicting the emergence of new linkages over time. In more formal terms, imagine we have a snapshot of a social network at current time t, and we want to reliably anticipate the edges that are not present at time t but are there at time t+h.

Similarity-based techniques assign a score to a pair of nodes that reflect a non-existent link. The number of common neighbors between them can be used to calculate similarity scores. This strategy has been investigated in a number of research works, with a positive association between link construction and greater common neighbor values. For each pair of nodes x and y, link prediction algorithms based on similarity provide them a similarity score. This score is based on how similar or close x and y are [18, 19]. For link prediction, maximum likelihood algorithms have just been presented in the recent past. By studying the structures and determining the highest likelihood, these algorithms are able to obtain specific rules and parameters. The criteria and parameters are used to anticipate new linkages. Even with a few thousand nodes, these methods are typically exceedingly sluggish. Various probability-based models have been pro-posed in the past. The Relational Bayesian Network, Probabilistic Entity-Relationship Models (PERMs), and Stochastic Relational Models are some of the examples of Probability Relational Models (PRMs) [24, 25]. Several writers have also proposed several nature-inspired link prediction techniques [10, 11, 12]. These methods mostly rely on physics rules and attempt to incorporate multiple net-work measurements into the formula. A gravity law-based link production model was proposed by Bastami, Mahabad, and Taghizadeh. For the reduction of the prediction space, they used community and network information. To replace the mass and distances in the gravitation law, Wahid-Ul-Ashraf, Budka, and Musial employed a variety of combinations for similarity and centrality of nodes. In Newton's Gravitation Law, they employed and evaluated a combination of three node centrality measurements as mass and thirteen dissimilarity values as distance. In the past, Node Embeddings have also been used for the task of link prediction [5, 6]. Node Embeddings are the low dimensional vector representation of nodes of the social network which can also be used in various machine learning algorithms. The advantage of using such techniques is that they abstract the underlying network.

Merger-based techniques tend to take up multiple similarity indices and devise a formula taking into account the effect of all these measures. These provide a performance boost since a single similarity measure works best on a limited type of network topologies. In the past fusion-based methods have been proposed including the Adaboost-based [27] and Ensemble-based Link Prediction algorithms [28]. We propose a method called Histogram-based Gradient Boosting Regression Tree for Link Prediction (HGBLP) based on the fusion of three popularity and four similarity measures based on the aforesaid study.

Proposed Method

Popularity measures used:

Degree Centrality. Degree centrality is the same as a node's degree in the con-text of graph theory. In other words, degree centrality is the number of nodes directly connected to the node in the network. This is a straightforward yet very helpful popularity index because in a way it precisely gives us local information about a network node’s popularity.

Closeness Centrality. This index measures a node's average distance i.e. how well connected on an average that particular node is to all other nodes in the network. The inverse of the sum of shortest paths from that node to all other nodes is used to calculate this value.

Page Rank. This is a centrality metric that is based on the network's random walk. Consider an imaginary traveler who is moving from one node to another picking his next node randomly. This centrality measures the likelihood that this traveler will arrive at a particular node in consideration.

Similarity measures used:

We also propose the use of the following five similarity measures in the algorithm:

Common Neighbors. This is a local similarity statistic based on the number of mutual friends shared by the node pair in question. Consider two nodes having a set of neighbors, the intersection of these sets represents the common neighbors.

Adamic Adar. This is an extended local similarity metric that is based on the number of mutual neighbors also taking into account the degrees of the nodes.

Cosine Similarity. The cosine similarity of two vectors represents the cosine of the angles between them. In network theory, if the nodes are represented as vectors and the cosine of the angle between them is taken, it forms an interesting similarity measure. It can be represented as Closeness centrality divided by the square root of the product of the degrees of the corresponding two nodes.

Matrix Forest Index. MFI [17] is defined as the ratio of the number of spanning forests such that v_i and v_j are contained in the same tree with v_i as the root to all spanning rooted forests of the main graph. The spanning subgraph may not have all of the graph's edges, but it does contain all of the graph's vertices.

Methodology

The steps involved in the proposed method are given below:

Reducing the network. An original network G(V, E) containing a set of vertices V and a set of edges E is used to generate a new network G'(V, E'). G' contains all of the vertices from the original graph G, and E' is a proper subset of E. This assures that some of the links in the newly established network G' are absent. We try to forecast the network G using the network G' for link prediction.

Computing the popularity measures. The newly established network is used to calculate the popularity measures for each node in the network. The popularity measures mentioned above are calculated and they will form the input for the regressor.

Computing the similarity indices. The similarity indices mentioned above are calculated and stored in a vector. The similarity measures are calculated for all the node pairs: (V × V) - E’. The node pairs can be classified into two categories, one of them representing the edges present in the original graph but not in the new graph representing G’ and the other category representing the edges that were never present in G. The task is essentially predicting the edges of the first category accurately.

Training and Prediction. The data created in steps 2 and 3 is down-sampled and split in a 7:3 ratio for training and testing. We apply Histogram Based Gradient Boosting Regression Tree regressor to train our model and then test it to predict the relationships.

Datasets used

The below datasets were used for the analysis of our proposed solution. These datasets are freely available, and the following is a description of them:

Jazz Musicians. Jazz performers represent the nodes, while an edge symbolizes two musicians who have collaborated [13].

Les Misérables. The characters in Les Misérables are represented by this data [13].

Football. The network of American Football games played amongst Division IA institutions during the regular season in Fall 2000 is shown by this data. [14]

Dolphin. Bottlenose dolphins form a social network by interacting with each other. The associations between dolphins are represented by the links in this data. [15]

US-Air. It has 332 airports and 2126 routes, with each node representing an air-port and each edge indicating whether or not a direct flight exists between two airports.

The data was preprocessed to provide a standard format with integer nodes. We didn't need to care for weights because the edges are unweighted. Table VI summarizes the common characteristics of the datasets mentioned previously.

TABLE VI. DATASETS USED FOR THE EVALUATION OF THE PROPOSED SOLUTION.

Dataset	Number of vertices	Number of edges	Average Degree
Jazz Musicians	198	2742	27.6970
Les Misérables	77	254	6.5974
Football	115	613	10.6609
Dolphin	62	158	5.0968
US-Air	332	2126	12

Evaluation Metrics

We used the area under the receiver operating characteristic (ROC) curve to assess the performance of our Histogram-based Gradient Boosting Regression Tree for Link Prediction (HGBLP) technique. AUC is a well-known metric that represents the likelihood of assigning a lower score to a randomly chosen non-existent connection than to a randomly given missing link. Let's say we do N separate comparisons of our algorithm's derived scores, and out of those N comparisons, missing connections get a higher score N' times. In addition, if the score is less than or equal N" times, the AUC value is defined by Equation (1).

$$AUC = \frac{N' + 0.5N''}{N} \tag{1}$$

Results

Table VI represents the calculated AUC values of our proposed Histogram-based Gradient Boosting Regression Tree for Link Prediction (HGBLP) method. The algorithm has shown a good performance boost in general on the datasets evaluated in the analysis.

TABLE VII. CALCULATED AUC VALUES FOR THE PROPOSED HISTOGRAM-BASED GRADIENT BOOSTING REGRESSION TREE FOR LINK PREDICTION (HGBLP) METHOD.

Dataset	Calculated AUC Value
Jazz Musicians	0.888650453
Les Misérables	0.919090909
Football	0.787076023
Dolphin	0.826190476
US-Air	0.898901617

Table VIII represents the calculated AUC values for the method proposed by the authors in literature [9] using the nature-inspired approach. In the literature, they suggested the use of the combination of Closeness centrality and Matrix Forest Index

(MFI) for best results. We used the method suggested by the authors using the same parameters for the calculation of the AUC values on the aforementioned datasets.

It is evident from the results that our proposed solution outperformed the nature-inspired approach on each dataset used. The results are compared in Fig. 9.

TABLE VIII. CALCULATED AUC VALUES USING THE NATURE INSPIRED APPROACH - CCM

Dataset	Calculated AUC Value
Jazz Musicians	0.522205232
Les Misérables	0.704840614
Football	0.566232998
Dolphin	0.630699088
US-Air	0.5862818

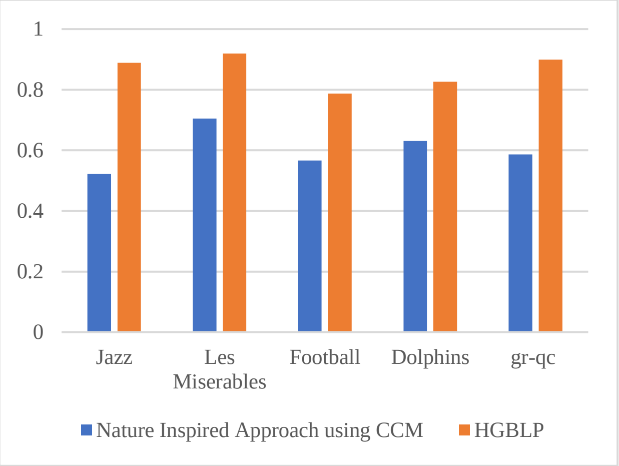

Fig. 9. Comparison of the AUC values on different datasets using the Nature Inspired Approach CCM and our proposed HGBLP solution.

We also compared and evaluated our solution against the Matrix Forest Index similarity measure [19]. The calculated AUC values for the Matrix Forest Index link prediction task are shown in Table IX. When compared with our solution, it is evident that our solution performed better on the used datasets on each instance (see Fig. 10).

TABLE IX. CALCULATED AUC VALUES USING MFI (MATRIX FOREST INDEX) FOR LINK PREDICTION.

Dataset	Calculated AUC Value
Jazz Musicians	0.506762996
Les Misérables	0.725754765
Football	0.570047309
Dolphin	0.569040382
US-Air	0.511013286

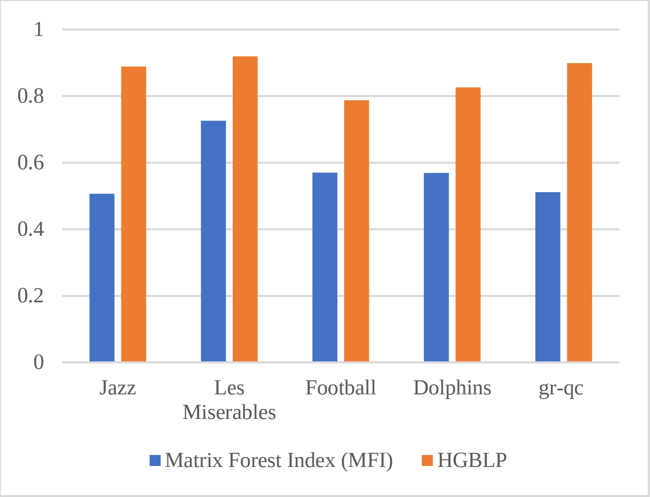

Fig. 10. Comparison of the AUC values on different datasets using the Matrix Forest Index (MFI) and our proposed HGBLP solution.

Finally, we evaluated our solution against the previously proposed fusion-based Ensemble Based Link Prediction (EMLP) algorithm defined in the literature [12]. The calculated AUC values for the link prediction task using the EMLP solution are shown in Table X.

TABLE X. CALCULATED AUC VALUES USING EMLP

Dataset	Calculated AUC Value
Jazz Musicians	0.864144785
Les Misérables	0.869090909

Football	0.747076023
Dolphin	0.726190476
US-Air	0.826321688

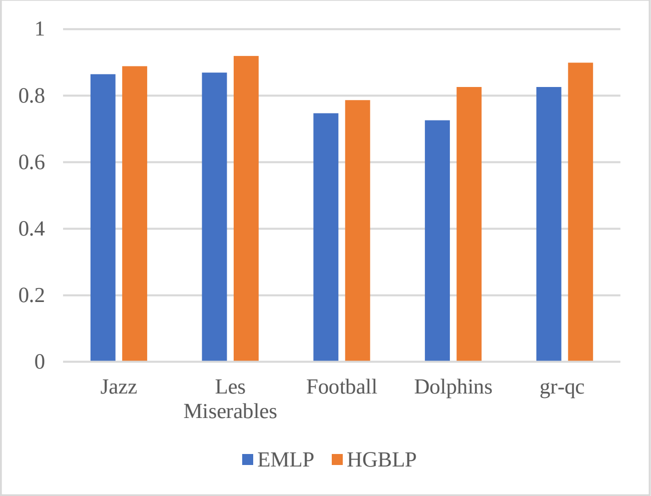

Fig. 11. Comparison of the AUC values on different datasets using the EMLP and our proposed HGBLP solution.

It is evident from Fig. 11 that our proposed solution HGBLP outperforms the previous fusion-based solution on the datasets mentioned in Table 1.

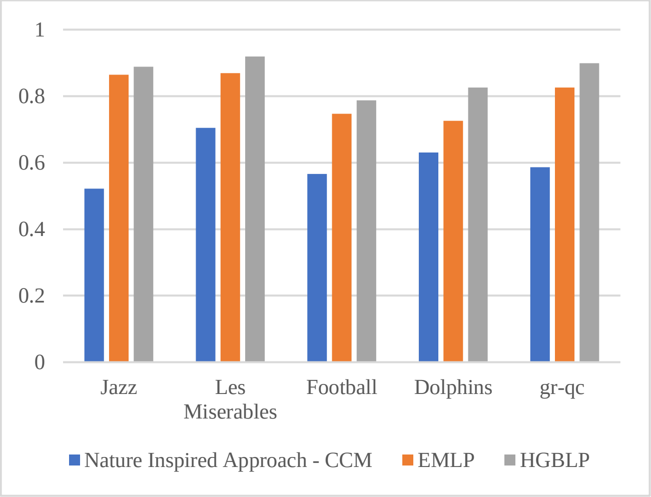

Fig. 12. Comparison of the AUC values on different datasets using all the fusion-based algorithms

The comparison of all the fusion-based solutions in Fig. 12 also suggests that the proposed HGBLP solution outperforms the previous fusion-based solutions. The nature-inspired approach fused one similarity and one popularity measure in the form of gravitational law formula.

CHAPTER 6 CONCLUSION

Based on the experiments conducted for the proposed work in this study, we can conclude that our nature-based solution outperforms the previous nature inspired approach as well as the individual algorithms used in the formulae (node2vec, MFI, and nature inspired - CCM), when used with cosine distances in node embeddings produced by node2vec instead of the proposed similarity/dissimilarity in the previous works. The future work for the study can explore the combination of different node embeddings in combination with different centrality measures.

The fusion-based solutions like the like EMLP and Nature-inspired solutions generally outperform the traditional single similarity-based solutions because they tend to average out the effect of a particular topology and structural property working in favor or against a particular measure. Finding the structural properties in an evolving social network is a challenging task and fusion-based solutions might be a good choice in such cases. Our proposed solution, HGBLP works better than the previously proposed methods in the literature.

REFERENCES

[1] Liben-nowell, David & Kleinberg, Jon. (2003). The Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology. 58. 10.1002/asi.20591.

[2] Kumar, S., Singhla, L., Jindal, K. et al. IM-ELPR: Influence maximization in social networks using label propagation based community structure. Appl Intell (2021). https://doi.org/10.1007/s10489-021-02266-w

[3] Lü, Linyuan, and Tao Zhou. "Link Prediction in Complex Networks: A Survey." Physica A: Statistical Mechanics and its Applications 390.6 (2011): 1150–1170. Crossref. Web.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119.

[5] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[6] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. 2016. Asymmetric Transitivity Preserving Graph Embedding. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1105–1114. DOI:https://doi.org/10.1145/2939672.2939751

[7] Katz, L. A new status index derived from sociometric analysis. Psychometrika 18, 39–43 (1953). https://doi.org/10.1007/BF02289026

[8] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural Deep Network Embedding. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1225–1234. DOI:https://doi.org/10.1145/2939672.2939753

[9] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," in Neural Computation, vol. 15, no. 6, pp. 1373-1396, 1 June 2003, DOI: 10.1162/089976603321780317.

[10] Bastami, E., Mahabad, and Taghizadeh. 2019. “: A gravitation-based link prediction approach in social networks. Swarm and evolutionary computation.” 176-186.

[11] Wahid-Ul-Ashraf, A., Budka, M., and Musial, K. 2019. “How to predict social relationships—Physics-inspired approach to link prediction.” Physica A: Statistical Mechanics and its Applications, no. 523, 1110-1129.

[12] Kumar S., Chaudhary U., Kedia R., Singhal T. (2021) Link Prediction in Complex Network: Nature Inspired Gravitation Force Approach. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_28

[13] “SNAP Datasets.” n.d. SNAP Datasets. https://snap.stanford.edu/data.

[14] “American College Football Dataset.” n.d. http://www.casos.cs.cmu.edu/computational_tools/datasets/external/football/index2.html.

[15] “Dolphins Dataset.” n.d. http://networkrepository.com/soc-dolphins.php.

[16] “GR-QC Dataset.” n.d. https://snap.stanford.edu/data/ca-GrQc.html.

[17] P. Chebotarev, E. Shamis, The matrix-forest theorem and measuring relations in small social groups (2006) arXiv preprint math/0602070.

[18] P. Holme, M. Huss, Role-similarity based functional prediction in networked systems: application to the yeast proteome, J. R. Soc. Interface 2 (2005) 327.

[19] G. Kossinets, Effects of missing data in social networks, Social Networks 28 (2006) 24.

[20] D. Lin, An information-theoretic definition of similarity, in Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufman Publishers, San Francisco, 1998.

[21] M. Sales-Pardo, R. Guimer`a, L. A. N. Amaral, Extracting the hierarchical organization of complex systems, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 15224.

[22] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 7821.

[23] Kumar, S., Lohia, D., Pratap, D. et al. MDER: modified degree with exclusion ratio algorithm for influence maximisation in social networks. Computing (2021). https://doi.org/10.1007/s00607-021-00960-8

[24] D. Heckerman, C. Meek, D. Koller, Probabilistic Entity-Relationship Models, PRMs, and Plate Models, In Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004, p. 55.

[25] K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu, Stochastic Relational Models for Discriminative Link Prediction, In Proceedings of Neural Information Precessing Systems, MIT Press, Cambridge MA, 2006, p. 1553.

[26] Kumar S., Kumar M. (2019) Predicting Customer Churn Using Artificial Neural Network. In: Macintyre J., Iliadis L., Maglogiannis I., Jayne C. (eds) Engineering Applications of Neural Networks. EANN 2019. Communications in Computer and Information Science, vol 1000. Springer, Cham. https://doi.org/10.1007/978-3-030-20257-6_25

[27] Z.F. WU , Q. Liang , Q. Liu , Z.G. Qin , Modified link prediction algorithm based on AdaBoost, J. Commun. 35 (2014) 116-123

[28] Li, K., Tu, L., & Chai, L. (2020). Ensemble-model-based link prediction of complex networks. Computer Networks, 166, 106978

LIST OF PUBLICATIONS OF THE CANDIDATE'S WORK

[1] Himanshu Tiwari, and Sanjay Kumar. "Link Prediction in Social Networks Using Node Embeddings with Nature Inspired Approach" In *3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–21)*. 2021.

[2] Himanshu Tiwari, and Sanjay Kumar. " Link Prediction in Social Networks using Histogram Based Gradient Boosting Regression Tree " *In IEEE 2021 International Conference on Smart Generation Computing, Communication and Networking (SMARTGEN).* 2021