Link Prediction in Online Social Networks

Sangeeta Kadam[1] Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, India s.kadam@ssipmt.com	Riju Bhattacharya[2] Associate Professor, Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, India riju@ssipmt.com	Dr J P Patra[3] Professor, Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, India patra.jyotiprakash@gmail.com

***Corresponding Author**: riju@ssipmt.com

Article History: Received: 05.10.2022 Revised: 23.12.2022 Accepted: 17.1.2023

ABSTRACT

Accurate link prediction across a large user base has become a challenging problem as online social networking platforms alter the ways and means of communication. Numerous applications, including friend recommendations, news commentary, and product recommendations, are affected by the issue. In this research, we provide a brand-new algorithm to address this issue. Due to their limitations in making full use of information or capturing all the features, the present online social network link prediction algorithms have various shortcomings in their link prediction accuracy. This study presents a novel formulation of the link prediction issue as a matrix denoising problem. We first suggest and thoroughly describe an unsupervised marginalized denoising model (USMDM). A mapping function that can find patterns in a massive amount of user data and comprehends the topological structure of social networks is the basis of the USMDM. A target matrix is projected onto the observed matrix via the mapping function. The initial matrix in the learning process is replaced with a low-rank matrix to increase effectiveness and avoid overfitting. The function can be trained on small datasets using the weak law of large numbers. Experiments are carried out on four actual social networks to show how well the suggested algorithm works, and the outcomes show how well the model works.

Keywords:* *Social networks, link prediction, matrix demolishing, weak law of large number.

I. INTRODUCTION

The online social network (OSN) is altering every area of people's life as businesses like Facebook and Twitter continue to expand their user bases. The OSN is created as a complement to actual in-person contacts with the aim of giving people another means of communication. The OSN's enormous potential as a man-made virtual network, however, gives its members the ability to reside entirely in a virtual social circle and create lasting relationships without having any face-to-face encounters. One of the core features of any OSN service, link prediction, is a popular area of study. There are many uses for an algorithm that can deliver precise link prediction results. If accurate, friend recommendations [1], [2], for instance, greatly enhance the user experience and quality of any online social network service [3]–[6]. As a result, it quickens the network's expansion [7]. In the business world, link prediction is also useful for finding trip companions on Dopplr, product recommendations on LinkedIn, and news feedback [8]. Link prediction in this study refers to the use of existing social network data to forecast new links. In Fig. 1, graph N displays known links as solid black

lines, while graph N displays unknown links (blue dotted lines). Based on the knowledge of graph N, the objective is to forecast unknown linkages of graph N.

Links in social networks have been predicted by several academics. Some researchers only focused on the similarity between nodal morphology because they were persuaded that relationships are likely to establish between two comparable people.

pairings [9]–[11], as well as others, adopted a different strategy. Many researchers have shifted their focus to the topology of social networks [12]-[14] as part of the ongoing effort to develop a method for precise link prediction because the structure and characteristics of the network are always crucial when determining whether two nodes form a link, regardless of the number of nodes and edge attributes within the network. Even if the study of topology has advanced significantly, some researchers continue to maintain alternative beliefs on link prediction, including the "social theories" [15], [16] like community theory, strong and weak connection theories, and homo genetic theory. Other academics have suggested learning-based strategies based on traits, intrinsic qualities, and extrinsic data [17, 18]. suggested the mDA method for domain adaptation, and the forecast outcomes appear to be on track. The method, however, appears to have two problems. Because mDA is only intended to capture network information on a "local" scale, it is unable to comprehend the entire spectrum of network information on a "global" scale and is therefore unable to utilise data on user characteristics. A novel link prediction model called USMDM has never been as effective as mDA. To create much more precise results, it fully utilises user attributes data as well as local and worldwide network topology data. Existing social networks are viewed by USMDM as being insufficient (an ideal social network is one that has investigated all potential links and connects its participants to the greatest extent; it is also called an ‘‘all links’’ network, where the term ‘‘

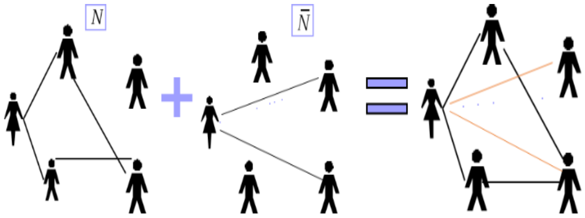


Figure 1: An illustrate example of link prediction.

In a network architecture known as "all links," links exist between nearly all user pairs (with the exception of a select few unique user pairings). The denoising auto-encoder technique used in deep learning is comparable to this concept. The initialization of the depth network in the denoising auto-encoder method frequently employs the denoising approach. Chen et al. [44] based on matrix denoising theory. The depth network's "coding" layer, which can better profile data with missing information, uses the "decoding" layer to rebuild the entire data set as its underlying process. The observed social network is the network with missing information and many links are not detected if the ideal social network is viewed as a complete data collection. Building a mapping function that can transform an imperfect social network into a perfect social network is the secret to the USMDM algorithm. Two elements affect the link-building process in social networks: similarity, such as a shared interest or educational background, and common acquaintances or friends. Consider the examples of users u1 and u2. A connection between users u1 and u2 can be predicted if they share a buddy, u3, or if u3 is friends with u2r's friend, u4. We treat this issue as a weight matrix minimization problem in order to generate the necessary function, and the result is the function of interest.In social network link's prediction, user's feature information and existing link information are very important for link prediction. But some algorithms, such as CN and AA, can only use a certain kind of information, but cannot make comprehensive use of it. The USMDM algorithm can make full use of these two kinds of information to predict, which can improve the accuracy of prediction. In the real world, for various reasons, people are unwilling to provide too much information, or even false information, which makes user features full of ‘noise’’. This will affect the prediction accuracy of the method using only user feature information. Althoughfeatures are also used in USMDM, the number of features needed can be relatively small. ‘‘Noise’’ features have littleinfluence on the algorithm, and features are only a part of the information required by the algorithm. The two

It is challenging to achieve both homophily and stochastic equivalence, two social network features. Some algorithms can only obtain one of these, while the USMDM algorithm can acquire both features simultaneously, increasing prediction accuracy. Additionally, the USMDM algorithm's computational complexity is greater than that of classical algorithms, which lowers its computational efficiency. Additionally, we will continue to research these flaws in the following years.

II. Literature Review

Numerous link prediction algorithms have been reported in recent years. These approaches can be divided into groups including similarity-based approaches, maximum likelihood approaches, and probabilistic model-based approaches.

Each node pair is given an index, which is defined as the similarity between the two nodes, in the similarity-based method. The similarities of all detected relationships are ranked, and the links connecting more similar nodes are thought to have greater existence likelihoods [6]. Numerous research, such as friendship prediction in [7], which examined the presence of homology in three systems that mix tagging social media with online social networks, discovered significant levels of topical similarity among users who are close to one another in the social network.

Based on maximum likelihood estimate, there is another class of link prediction techniques. These techniques assume a few network structure organising models with particular parameters and precise rules that are obtained by maximising the likelihood of the observed structure. Then, using those guidelines and parameters, it is possible to determine the likelihood of any unobserved relationship. The hierarchical structure model and the stochastic block model are typical network organising models [8–11]. In [12], a set of straightforward features is put forth as a structural model that can be examined to spot gaps. A hierarchical model can handle a network of important organizational levels with high precision, such as a terrorist network or a network of grasslands' food chains. However, due to the high computational complexity required to handle large scale networks, it must generate a huge number of samples in order to predict the network.

The probability model serves as the foundation for another kind of link prediction tool. By removing the underlying structure from the observed network, these model-based approaches seek to anticipate the missing links by applying the learnt model. In order for the resulting model to have better structures and relationships reflecting actual network features, these approaches first construct a model with a set of changeable parameters, and then they utilise an optimization strategy to determine the ideal parameter values.

Reference [13] demonstrated that by identifying the evolutionary model of triads between two consecutive snapshots of a network, new linkages can be inferred. A supervised structural link prediction algorithm is what [13] proposes. There would be 64 distinct triads in a directed graph. As shown in Fig. 1, there would be four different types of connections—one two-way connection, two one-way connections, and one no-connection—between every pair of nodes in a triad. This network's 64 separate triads can be counted in two continuous snapshots to create a matrix known as the Triad Transition Matrix (TTM). The likelihood of a link forming between two unconnected nodes can be calculated using the TTM matrix.

The [14] writers algorithm for structural supervised link analysis and prediction. Vertex Collocation Profiles are substructures of a graph that are discovered using the technique (VCP). This algorithm can be used for link prediction if a learning phase is added to it. This algorithm's disadvantage is that it takes a lot of time and is impractical for VCPs in big networks with more than four nodes. Special subgraphs in directed networks were examined by Zhang and his colleagues [15], who dubbed them the microscopic organising principles of directed networks. Some of these subparagraphs are more prevalent in social networks, according to their studies. The Bi-fan structure, which has 4 nodes and 4 directed links, is the most popular local structure in directed networks. According to homophily, they have demonstrated this theory [16].and clustering mechanism and potential theory. Subgraphs that have only one link fewer than Bi-fan structure, that link has the highest probability to be created in the near future. This is the principal idea of the link prediction algorithm introduced in [15].

A concept known as supervised random walks was created by Leskovec et al. [17]. It creates a unified link prediction method by fusing the network structure with the characteristics of the nodes and edges of the network. After that, they create a strategy based on it. In a supervised manner, the approach learns to separate a PageRank-like random walk over the network, making it more likely to visit nodes to which new linkages will be created in the future. A model including theories of balance and status from social psychology is used to predict the signals of connections in social networks [18]. Relationships can be either good (friendship) or negative (opposition) in social networks. Two types of features are utilised to integrate the analysis of signed networks with machine learning methods. One is based on the number of nodes, while the other is based on a social psychology theory. They also look at the issue of incomplete networks, when nodes and edges are both absent. In order to estimate the missing portion of the network, they additionally construct KronEM, an EM technique coupled with the Kronecker graphs model [19]. Leskovec et al. also assembled a large number of social network datasets that are accessible to other researchers. These datasets have been incorporated into numerous link prediction studies.

In directed social networks, Hopcroft and Tang's team [20] investigates the novel challenge of reciprocal connection prediction to foretell who will follow you back. In order to include social theories (such as structural balance and homophily) over triads into the semi-supervised machine learning model, they introduced the Triad Factor Graph (TriFG) model. A prediction problem was also developed by Tang's team [21] to foretell the existence and nature of linkages between a pair of nodes. To capture the interrelationship influence, they introduced a partially-labeled pairwise factor graph model (PLP-FGM) and two active learning procedures (Influence-Maximization Selection and Belief Maximization Selection) [22]. They expanded the aforementioned approach to address the issue of inferring social relationships across diverse networks [23]. The model incorporates social theories into a framework for semi-supervised learning that allows supervised data to be transferred from a source network to a target network in order to infer social links there. For the inventor social network, where convention connections serve as the link between inventors. In order to recommend patent partners, they also include user interactions into a factor graph model [24]. This approach has good prediction accuracy and efficiency, therefore it might help existing user-feedback-based recommendation models.

III. The Traditional Link Prediction Algorithms

An important area of data mining study is link prediction. There are many different scenarios in it. The relationships between the objects are often considered in data mining jobs. Recommendation engines, social networks, information retrieval, and many more domains can all benefit from link prediction.

Link prediction is the process of estimating the likelihood that the nodes Vi and Vj will link together given a snapshot graph of the social network at the time G = V, E and the nodes Vi and Vj. The definition of link prediction demonstrates the division of the link prediction problem into two categories. Predicting when the new link will emerge falls under the first category. The second category entails predicting a space's hidden, undiscovered links.

Based on the algorithm's similarity, the link prediction framework is the simplest. Any pair of nodes x and y that we have assigned to this node is a function called Similarity (x, y), which is defined as the function that measures how similar the two nodes are to one another. The chance of a link between any two nodes increases with the value of the similarity function, which is sorted from biggest to smallest among the pairs of nodes.

Here, we present a few straightforward similarity indices for link predictions.

3.1. Local Similarity Index

3.1.1. Common Neighbors. Assume that the node V $\in V$; then the neighbors of the node set $\Gamma(\mathrm{V}) := \{t \mid (t, \mathrm{V}) \in E \vee (\mathrm{V}, t) \in E \wedge t \neq \mathrm{V}\}$; that is, $\Gamma(\mathrm{V})$ is the set of all the neighbors of node V. The common neighbors of node u and node V refer to the jointly owned neighbors by node u and node V For the undirected graph, the common neighbors can use the following definition:

Similarity $(u, \mathrm{V}) = |\Gamma\ (u) \cap \Gamma\ (\mathrm{V})|$. (1)

Kossinets and Watts analysis of the large-scale social network, found that two students who have more mutual friends will have greater possibility to become friends [7].

3.1.2. Preferential Attachment.

Preferential attachment mechanism can be used to generate scale-free network evolution model. The probability of generating a new link of node u is directly proportional to the degree of the node [8]. This is the same as the truth “the rich are getting richer” in economics. Therefore, the probability of the link between node u and node V is directly proportional to $d_u \times d_{\mathrm{V}}$. Inspired by this mechanism, the PA similarity index can be defined as follows:

Similarity $(u, \mathrm{V}) = du \times d\mathrm{V}$. (2)

It may be noted that the similarity index does not require any node neighbor information; therefore, this similarity index has the lowest computational complexity.

3.1.3. Adamic-Adar [9]. This similarity index assigns a higher similarity function value to a small degree node. AdamicAdar algorithm believes that an affair owned by less objects, compared to owned by more objects, has greater effect on link prediction. Its definition is as follows:

Similarity $(u, \mathrm{V}) = \sum_{\mathrm{z} \in \Gamma(\mathrm{u}) \cap \Gamma(\mathrm{V})} \frac{1}{\log \mathrm{dz}}$.(3)

3.1.4. Resource Allocation. This similarity index is inspired by the ideas of complex network resources dynamically allocated [10]. In pair of nodes u, V that have no direct link, node u can allocate some resources to the node V through their common neighbor. Their common neighbors assume the role of passers. In the simplest case, we assume that each passer has a unit of resources; it assigns these resources to its neighbors evenly. Therefore, the similarity of node u and node V can be defined as the number of resources that node u get from node V; namely,

Similarity $(u, \mathrm{V}) = \sum_{\mathrm{z} \in \Gamma(\mathrm{u}) \cap \Gamma(\mathrm{V})} \frac{1}{\mathrm{dz}}$. (4)

3.2. Overall Similarity Index

3.2.1. Katz [11]. In 1953, Katz described the similarity using the global path. The idea of the method is that the more paths between two nodes are, the greater the similarity between two nodes is. Katz measure is defined as follows:

Similarity $(u, \mathrm{V}) = \sum_{i=1}^{imax=\infty} \beta l \cdot |\mathrm{path} l\ \mathrm{u,v}|$ (5)

$= \beta A_{u\mathrm{V}} + \beta^2 (A^2)_{u\mathrm{V}} + \beta^3 (A^3)_{u\mathrm{V}} + \cdots$,

where $|\mathrm{path}^{l}{}_{u,\mathrm{V}}|$ is the number of paths between node u and node V and the length of the path is l. β is a parameter between 0 and 1. This parameter is used to control the contribution of path to the similarity; the longer the path is, the less contribution the path made to the similarity. The computational complexity of Katz measure is n^3 , so the measure is not suitable for large-scale network.

3.2.2. Random Walk with Restart (RWR) [12]. This indicator is a direct application of the PageRank algorithm. A random walker starting from node u will reach its random neighbor with probability c repeatedly and return the node u with the probability $1-c$. $q_{u\mathrm{V}}$ represents the probability of the random walker reaching node V in the steady state condition. Therefore, we have $\vec{q}_u = cP^T\vec{q}_u + (1 - c)\vec{e}_u$, where P is the transfer matrix. If the node u is connected with node V, then $P_{u\mathrm{V}} = 1/d(u)$; else $P_{u\mathrm{V}} = 0$. So the solution is simple; namely, $\vec{q}_u = (1 - c)(I - cP^T)^{-1}\vec{e}_u$. RWR coefficient can be defined as

Similarity $(u, \mathrm{V}) = q_{u\mathrm{V}} + q_{\mathrm{V}u}$. (6)

Compared to the local similarity index, the global similarity index needs more overall network topology information. Although the performance of the overall similarity index is better than the local similarity index, it has two fatal flaws: first, the global similarity index calculation is very time consuming, and when the network is huge, this calculation program of the global similarity index does not work. Second, sometimes the global topology information is not available, especially when we use a decentralized approach to implement the algorithm. Therefore, how to design a similarity index is particularly important, which is easy to calculate and its accuracy is high. Although the traditional link prediction algorithms have made some prediction effect, they do not make full use of the topology information. Common neighbor algorithm treats all the common neighbors equally; it does not distinguish the different neighbors' different effects on the link prediction. Katz algorithm distinguishes the different path's influences which have different lengths, but it does not distinguish the influence of the paths with the same length on link prediction. These algorithms only consider the topology characteristics of the network, treat the social networking static, and ignore the time attributes and node attribute of social network. How to integrate the topology characteristics, time characteristics, and node attributes of social network reasonably is an enormous challenge for link prediction facing.

IV. Proposed User Behavior Based Link Prediction Method

Existing social network platforms provide features such as sharing, commenting, likingthe content shared by other users. These attributes can be used along with the existing attributes to improve the link prediction accuracy (Li et al., 2016). The below proposedmetrics are extensively used in online advertising or digital marking by companies like Google through AdSense program and Facebook through Facebook Ads (Measured, 2016). They use these metric to maximize the advertisement reach and only allow legitimate spending of customer budget. Using these metrics in link prediction problemin addition to topological structure of the social networks is a novel approach.

a) User Action Metrics

Three types of User action metrics are defined, namely Engagement rate, Reach Rate and Impression rate.

b) Engagement Rate

It measures users interaction with the post and promotion of the post to others circleof friends. It is a key metric for discovering how people engage with the post by shar- ing, commenting, likening or by clicking the post. The amount of engagement the postreceives makes it feasible to understand the nodes interest in the subject and similar nodes can be recommended for forming links. Relatively these are considered as po- tential users for recommendation.

Engagement Rate (ER) is calculated as the sum of likes, comments and shares madeby a node,

$$ER = |likes|+|comments|+|shares|$$

where $|likes|$ is the number of pages liked by the node (i.e., user), $|comments|$ is the number of comments made by the user and $|shares|$ is the number of posts or content shared by the user.

c) Reach Rate

Reach is also another key metric that social media marketers use to in any product or brand awareness. It is more accurate measure than page likes. Since all the people wholike the page may not see the posts and many users who do see the post may not like the page. Total reach rate is calculated from organic reach, paid reach and viral reach. Organic reach is defined as the number of unique people who saw the post in news feed.Paid reach is defined as the number of unique people who saw the post from Ads or thesuggested posts. Viral reach is defined as the number of unique people who saw the post published by a friend. For example, if a fan likes, comments or share the post, their friends see the post even if they are not fans of the page. If total reach rate increases numbers of Ad click also increases which in turn generates revenue to the facebook.

Total Reach Rate TR is calculated as the sum of Organic reach, Paid reach and viralreach.

$$TR = |organic\ reach|+|paid\ reach|+|viral\ reach|$$

d) Impression Rate

Impression rate is a key metric to understand how frequently users are exposed to the post content. Impressions are calculated by counting number of times the content associated with the page is displayed. Total impression rate is calculated from organic impression, paid impression and viral impression. Organic Impressions is defined as number of times the content was displayed in a users ticker, news feed. Paid impressionis defined as number of times the content was displayed through Ads. Viral Impressionis defined as number of times content associated with the page was displayed directed by a friend by liking, commenting and sharing. Total Impression rate TI is calculated as the sum of Organic Impression, Paid Im- pression and Viral Impression.

$$TI = |organic\ impression| + |paid\ impression| + |viral\ impression|$$

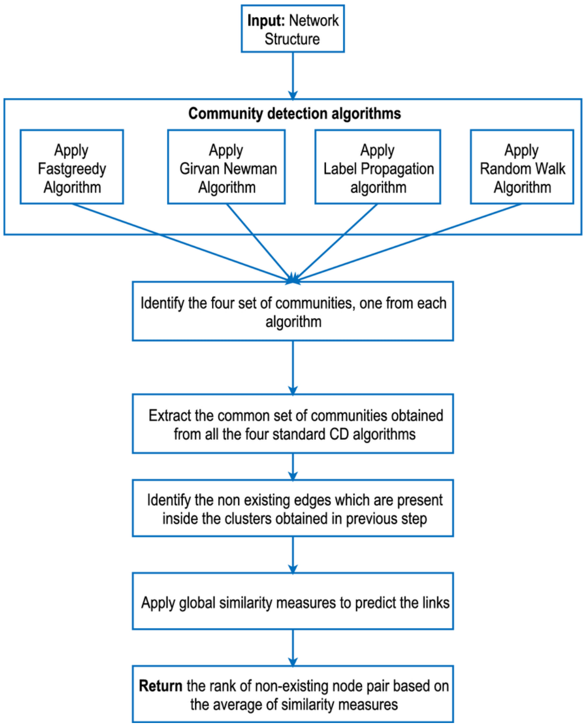
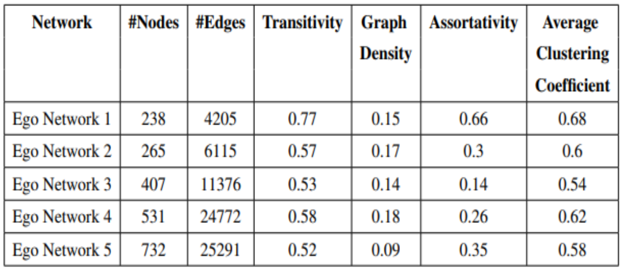

Input: Network Structure
Community detection algorithms
Apply Fastgreedy Algorithm
Apply Girvan Newman Algorithm
Apply Label Propagation algorithm
Apply Random Walk Algorithm
Identify the four set of communities, one from each algorithm
Extract the common set of communities obtained from all the four standard CD algorithms
Identify the non existing edges which are present inside the clusters obtained in previous step
Apply global similarity measures to predict the links
Return the rank of non-existing node pair based on the average of similarity measures

Fig2: Process flow Diagram

V. EXPERIMENTAL DATASET AND TOPOLOGICAL PROPERTIES

This work considers a real-time dataset of Facebook website gathered from the SNAP (Stanford Network Analysis Platform) [37] for the analysis of proposed approach, FIXT. The network dataset obtained from the SNAP repository consists of a total of 27,519 nodes (or users). The collected dataset also contains two additional details: Connectivity information of each node (or user) and, individual node's profile features (such as hometown, current city, etc.). The dataset contains a total of 1,143 ego networks. Five distinct ego networks having different sizes are arbitrarily considered for experimental evaluation. The ego networks for experimental evaluation are selected in such a manner that topological properties of the considered ego networks vary from each other. SNAP dataset has been widely used in literature for evaluation purposes [158] [159]. Table 5.1 presents the important topological properties for each of the five ego networks used in this work for experimental evaluation. The important topological properties considered are: node count, edge count, network transitivity, network density, network assortativity and average clustering coefficient. These important topological properties of a network help to understand the performance of various link prediction techniques in different network scenarios. The authors detail each topological property represented in Table 5.1 in detail as follows:

Table 1: Topological properties of network dataset used for experimental evaluation of proposed approach, FIXT.



Network	#Nodes	#Edges	Transitivity	Graph Density	Assortativity	Average Clustering Coefficient
Ego Network 1	238	4205	0.77	0.15	0.66	0.68
Ego Network 2	265	6115	0.57	0.17	0.3	0.6
Ego Network 3	407	11376	0.53	0.14	0.14	0.54
Ego Network 4	531	24772	0.58	0.18	0.26	0.62
Ego Network 5	732	25291	0.52	0.09	0.35	0.58

VI. Result:

Proposed method for link prediction in online social network exploiting the node neighborhood property is applied upon Facebook, Deezer, Github and Twitch online social networks. The probabilistic similarity measure computed contributes to the probability of having a link between the nodes across the social networks. The similarity between the nodes is identified by using the proposed similarity algorithm can be referred as probabilistic similarity measure described in the above section. The obtained probability similarity measures are tabulated in Table 2.The threshold value for the probability contribution '*p*' is set on the range of 0 to 1. Higher the probability contribution more will be the probability of having link between the node pair.

Table 2. Similarity obtained between the node pairs for future link prediction

Node u	Node v	Prob_similarity
4	56	0
4	57	0
4	58	0
4	59	0.01
4	60	0
4	61	0.01
4	62	0
4	63	0.01
4	64	0.01
4	65	0.01
4	66	0
4	67	0.01

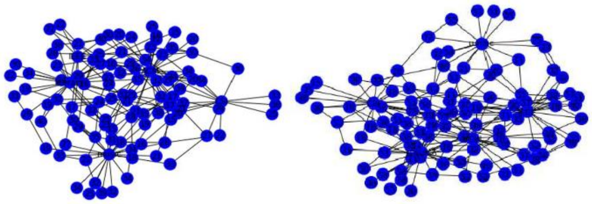


Figure 3. (a) Evolution of social network at time't'; (b) The graph obtained at time't+1'

the comparative analysis of proposed similarity algorithm on Tweet online social network. Similarly, Table 3, Table 4 and Table 5 are demonstration of Facebook, Deezer and GitHub online social networks respectively.

Table 3. Performance analysis of algorithm on Twitch online social network

Method	Precision	Recall	F-measure

AA	0.975	0.985	0.985
JC	0.969	0.979	0.979
PA	0.954	0.951	0.948
RA	0.98	0.981	0.981
Proposed Algorithm	0.982	0.981	0.982

Table 4. Performance analysis of algorithm on Facebook online social network

Method	Precision	Recall	F-measure
AA	0.956	0.953	0.95
JC	1	1	1
PA	0.959	0.957	0.953
RA	0.976	0.975	0.974
Proposed Algorithm	0.985	0.98	0.981

Table 5. Performance analysis of algorithm on Deezer online social network

Method	Precision	Recall	F-measure
AA	0.943	0.938	0.935
JC	1	1	1
PA	0.982	0.982	0.982
RA	0.953	0.95	0.949
Proposed algorithm	0.989	0.989	0.987

Table 4.6 Performance analysis of algorithm on Github online social network

Method	Precision	Recall	F-measure
AA	0.946	0.944	0.945
JC	1	1	1
PA	0.98	0.977	0.978
RA	0.979	0.979	0.979
Proposed Algorithm	0.985	0.982	0.98

VII. Conclusion

Social Networks model the interaction among the people and entities involved. Graph based analysis of such social networks provides rich information about the nature and evolution of network. Link prediction problem addresses this problem of how likely the entities involved are likely to form a connection over a period of time. Majority of the methods proposed in the literature analyze the link prediction problem only based on graph topology. In this thesis, the problem of link prediction in social networks is discussed considering additional features of the network nodes (i.e., users) apart from the graph topology. Social Networks comprise of large nodes and inherent topology that have to be analyzed to obtain meaningful inferences on link prediction. It is a difficult task for large graphs as traversing such large graphs require huge computational effort and optimized techniques. To address this issue, firefly optimization technique is employed. The fireflies are made to traverse on graph edge based on the brightness factor and the fireflies will be attracted to nodes that have a higher probability of forming links. Three variants of the firefly link prediction algorithm are proposed concentrating on structural link or topology of the network and on the attributes of nodes on which fireflies traverse. Through experiments, it was confirmed that the proposed algorithms performs better than the existing ones in the literature.

Future Scope:

Some potential areas of future development for link prediction in online social networks include:

Deep Learning Techniques: With the rise of deep learning techniques, there is potential for developing more accurate and robust link prediction models. Deep learning models can automatically learn feature representations of nodes and edges, which can be used to predict missing links in a network.

Incorporation of Temporal Information: Temporal information plays a crucial role in predicting links in online social networks. There is a need to develop models that can capture the temporal dynamics of social networks and predict future links based on historical data.

Integration of Heterogeneous Data: Online social networks contain a wide range of heterogeneous data, such as user profiles, posts, comments, and likes. Integrating this data with link prediction models can lead to more accurate predictions and better understanding of social network dynamics.

Evaluation of Robustness: Social networks are prone to attack, and link prediction models should be evaluated for their robustness to adversarial attacks. Developing models that are robust to attacks can help prevent the spread of misinformation and malicious content on social networks.

Cross-Domain Link Prediction: Link prediction can be applied to different domains such as e-commerce, transportation, and healthcare. Developing models that can predict links across different domains can lead to better understanding of complex systems and lead to improved decision-making.

In summary, link prediction in online social networks has a bright future with potential advancements in deep learning techniques, temporal modeling, heterogeneous data integration, robustness evaluation, and cross-domain link prediction.

REFERENCES

[1] P. Wang, B. Xu, Y. Wu, and X. Zhou, ‘‘Link prediction in social net- works: The state-of-the-art,’’ *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, Jan. 2015. doi: 10.1007/s11432-014-5237-y.

[2] W. Yu and G. Lin, ‘‘Social circle-based algorithm for friend recommendation in online social networks,’’ *Chin. J. Comput.*, vol. 37, no. 4,pp. 801–808, 2014.

[3] J. Dimicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Müller ‘‘Motivations for social networking at work,’’ in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, San Diego, CA, USA, 2008, pp. 711–720.

[4] J. Tang, S. Wu, J. Sun, and H. Su, ‘‘Cross-domain collaboration recommendation,’’ in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, Aug. 2012, pp. 1285–1293.

[5] M. Pavlov and R. Ichise, ‘‘Finding experts by link prediction in coauthorship networks,’’ in *Proc. 2nd Int. Conf. Finding Experts Web with Semantics (FEWS)*, Busan, Korea, 2007, pp. 42–55.

[6] T. Wohlfarth and R. Ichise, ‘‘Semantic and event-based approach for link prediction,’’ in *Proc. Int. Conf. Practical Aspects Knowl. Manage.*, 2008,pp. 50–61.

[7] Z. Yin, M. Gupta, T. Weninger, and J. Han, ‘‘LINKREC: A unified framework for link recommendation with user attributes and graph structure,’’in *Proc. 19th Int. Conf. World Wide Web*, New York, NY, USA, Apr. 2010, pp. 1211–1212.

[8] I. Guy, I. Ronen, and E. Wilcox, ‘‘Do you know?: Recommending people to invite into your social network,’’ in *Proc. 14th Int. Conf. Intell. Interfaces.*Sanibel Island, FL, USA, Feb. 2009, pp. 77–86.

[9] P. Bhattacharyya, A. Garg, and S. F. Wu, ‘‘Analysis of user keyword similarity in online social networks,’’ *Social Netw. Anal. Mining*, vol. 1,no. 3, pp. 143–158, Jul. 2011.

[10] C. G. Akcora, B. Carminati, and E. Ferrari, ‘‘User similarities on social networks,’’ *Social Netw. Anal. Mining*, vol. 3, no. 3, pp. 475–495, Sep. 2013.

[11] A. Anderson, D. Huttenlocher, and J. Kleinberg, ‘‘Effects of user similarityin social media,’’ in*Proc. 5th ACM Int. Conf. Web Search Data Mining*, Seattle, WC, USA, Feb. 2012, pp. 703–712.

[12] L. A. Adamic and E. Adar, ‘‘Friend and neighbors on the Web,’’ *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

[13] R. N. Lichtenwalter and N. V. Chawla, ‘‘Vertex collocation profiles: Sub-graph counting for link analysis and prediction,’’ in *Proc. 21st Int. Conf. World Wide Web*, Lyon, France, Apr. 2012, pp. 1019–1028

[14] P. Symeonidis and N. Mantas, ‘‘Spectral clustering for link predictionin social networks with positive and negative links,’’ *Social Netw. Anal. Mining*, vol. 3, no. 4, pp. 1433–1447, Dec. 2013.

[15] J. Valverde-Rebaza and A. de Andrade Lopes, ‘‘Exploiting behaviors of communities of twitter users for link prediction,’’ *Social Netw. Anal. Mining*, vol. 3, no. 4, pp. 1063–1074, Dec. 2013.

[16] H. Liu *et al.*, ‘‘Hidden link prediction based on node centrality and weakties,’’ *Europhys. Lett.*, vol. 101, Jan. 2013, Art. no. 18004.

[17] J. Zhu, ‘‘Max-margin nonparametric latent feature models for link predic-tion,’’ in *Proc. ICML*, Feb. 2012, pp. 719–726.

[18] J. R. Lloyd, P. Orbanz, Z. Ghahramani, and D. M. Roy, ‘‘Random function priors for exchangeable arrays with applications to graphs and relational data,’’ in *Proc. NIPS*, 2012, pp. 1007–1015.

[19] M. E. J. Newman, ‘‘Clustering and preferential attachment in growing networks,’’ *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip.Top.*, vol. 64, Aug. 2001, Art. no. 025102.

[20] L. A. Adamic and E. Adar, ‘‘Friends and neighbors on the Web,’’ *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

[21] J. Scripps, P.-N. Tan, F. Chen, and A. Esfahanian, ‘‘A matrix alignment approach for link prediction,’’ in *Proc. 19th Int. Conf. Pattern Recognit. (ICPR)*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[22] J. Kuncgis and A. Lommatzsch, ‘‘Learning spectral graph transformations for link prediction,’’ in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Montreal,QC, Canada, Jun. 2009, pp. 561–568.

[23] P. Symeonidis, N. Iakovidou, N. Mantas, and Y. Manolopoulos, ‘‘From biological to social networks: Link prediction based on multi-way spectralclustering,’’ *Data Knowl. Eng.*, vol. 87, no. 4, pp. 226–242, Sep. 2013.

[24] S. Bastani, A. K. Jafarabad, and M. H. F. Zarandi, ‘‘Fuzzy models for link prediction in social networks,’’ *Int. J. Intell. Syst.*, vol. 28, no. 8,pp. 768–786, Aug. 2013.

[25] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, ‘‘An evolutionary algorithm approach to link prediction in dynamic social networks,’’ *J. Comput. Sci.*, vol. 5, no. 5, pp. 750–764, Sep. 2014.

[26] A. E. Bayrak and F. Polat, ‘‘Contextual feature analysis to improve link pre- diction for location based social networks,’’ in *Proc. SNAKDD 8th Work- shop Social Netw. Mining Anal.*, Aug. 2014, Article No.7.

[27] X. Xie, Y. Li, Z. Zhang, S. Han, and H. Pan, ‘‘A joint link prediction method for social network,’’ in *Intelligent Computation in Big Data*. Herbin, China: ICYCSEE, 2015, pp. 56–64.

[28] H. Kashima and N. A. Abe, ‘‘A parameterized probabilistic model of network evolution for supervised link prediction,’’ in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Hong Kong, Dec. 2006, pp. 340–349.

[29] F. Fouss, A. Pirotte, J. Renders, and M. Saerens,, ‘‘Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[30] V. Leroy, B. B. Cambazoglu, and F. Bonchi, ‘‘Cold start link prediction,’’ in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Washington, DC, USA, Jul. 2010, pp. 393–402.

[31] J. Zhu, Q. Xie, and E. J. Chin, ‘‘A Hybrid time-series link prediction framework for large social network,’’ in *Proc. 23rd Database Expert Syst. Appl. (DEXA)*, 2012, pp. 345–359.

[32] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, ‘‘Deep learning approaches for link prediction in social network services,’’ in *Proc. 20th Int. Conf. Neural Inf. Process. (ICONIP)*, 2013, pp. 425–432.

[33] N. Gong *et al.*, ‘‘Joint link prediction and attribute inference using a social-attribute network,’’ *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 2, pp. 529–544, Apr. 2014.

[34] A. P. S. Panwar and R. Niyogi, ‘‘A heuristic for link prediction in online social network,’’ in *Intelligent Distributed Computing*. Berlin, Germany: Springer, 2015, pp. 31–41.

[35] Y.-L. He, J. N. K. Liu, Y.-X. Hu, and X.-Z. Wang, ‘‘OWA operator based link prediction ensemble for social network,’’ *Expert Syst. Appl.*, vol. 42, no. 1, pp. 21–50, Jan. 2015.

[36] E. Sherkat, M. Rahgozar, and M. Asadpourb, ‘‘Structural link prediction based on ant colony approach in social networks,’’ *Phys. A Stat. Mech. Appl.*, vol. 419, Feb. 2015, pp. 80–94.

[37] A.-T. Nguyen-Thi, P. Q. Nguyen, T. D. Ngo, and T.-A. Nguyen-Hoang, ‘‘Transfer adaboost SVM for link prediction in newly signed social net- works using explicit and PNR features,’’ *Procedia Comput. Sci.*, vol. 60, pp. 332–341, Jan. 2015.

[38] B. Zhu and Y. Xia, ‘‘Link prediction in weighted networks: A weighted mutual information model,’’ *PLoS ONE*, vol. 11, no. 2, Feb. 2016, Art. no. e0148265.

[39] D. Li *et al.*, ‘‘Link prediction in social networks based on hypergraph,’’, ACM, New York, NY, USA, Tech. Rep., 2016, pp. 41–42.

[40] J. Wu, G. Zhang, and Y. Ren, ‘‘A balanced modularity maximization link prediction model in social networks,’’ *Inf. Process. Manage.*, vol. 53, no. 1, pp. 295–307, Jan. 2017.

[41] E. Bastami, A. Mahabadi, and E. Taghizadeh, ‘‘A gravitation-based link prediction approach in social networks,’’ *Swarm Evol. Comput.*, vol. 44, pp. 176–186, Feb. 2019.

[42] Y. Sara and R. M. Mohammad, ‘‘A link prediction method based on learning automata in social networks,’’ *J. Comput. Robot.*, vol. 11, no. 1, pp. 43–55, Mar. 2018.

[43] D. P. Hoff, ‘‘Modeling homophily and stochastic equivalence in symmetric relational data,’’ in *Proc. NIPS*, 2007, pp. 1–8.

[44] M. Chen *et al.*, ‘‘Marginalized denoising autoenco-ders for domain adap- tation,’’ in *Proc. Int. Conf. Mach. Learn.*, Jun. 2012, pp. 1–8.

[45] J. McAuley and J. Leskovec, ‘‘Learning to discover social circles in ego networks,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 539–547.

[46] *Social Networks*. Accessed: Feb. 5, 2017. [Online]. Available: http://snap.stanford.edu/data/

[47] H. Li, Z. Bu, A. Li, Z. Liu, and Y. Shi, ‘‘Fast

[48] and accurate mining the community structure: Integrating center locating and membership opti- mization,'' *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2349–2362,Sep. 2016.

[49] Z. Bu, H. Li, J. Cao, Z. Wang, and G. Gao, ''Dynamic cluster formation game for attributed graph clustering,'' *IEEE Trans. Cybern.*, vol. 49, no. 1,pp. 328–341, Jan. 2019.

[50] H.-J. Li, Z. Bu, Z. Wang, J. Cao, and Y. Shi, ''Enhance the performance of network computation by a tunable weighting strategy,'' *IEEE Trans.Emerg. Topics Comput.*, vol. 2, no. 3, pp. 214–223, Jun. 2018.

[51] H.-J. Li and J. J. Daniels, ''Social significance of community structure: Statistical view,'' *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip.Top.*, vol. 91, no. 1, Jan. 2015, Art. no. 012801.