A NOVEL METHOD FOR LINK PREDICTION IN SOCIAL NETWORK

A Thesis

Submitted in partial fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science Engineering

By

Praveen Kumar Bhanodia

Reg. No. 41600126

Supervised By	Co-Supervised by
Dr. Aditya Khamparia	**Dr. Babita Pandey**

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
2020

Declaration

I hereby declare that the thesis entitled “A NOVEL METHOD FOR LINK PREDICTION IN SOCIAL NETWORK” submitted by me for the Degree of Doctor of Philosophy in Computer Science and Engineering is the result of my original and independent research work carried out under the guidance of Supervisor Dr. Aditya Khamparia and co-supervisor Dr. Babita Pandey, and it has not been submitted for to any university or institute for the award of any degree or diploma.

Place

Date Signature of the Student

Certificate

This is to certify that the thesis entitled “A NOVEL METHOD FOR LINK PREDICTION IN SOCIAL NETWORK” submitted by Praveen Kumar Bhanodia for the award of the degree of the Doctor of Philosophy in Computer Science and Engineering, Lovely Professional University Punjab, India is entirely based on the work carried out by him under our supervision and guidance, The work recorded, embodies the original work of the candidate and has not been submitted for the award of any degree, or diploma of any university and institute, according to the best of our knowledge.

Date   Signature of Supervisor

Signature of Co-Supervisor

Abstract

This research work has been established on the link prediction issues in online social network analysis discussed by Liben-Nowell and Kleinberg. Link prediction problem is an instance of online social network analysis where links are supposed to be predicted between the unconnected node pair across the social network in future based on the structural topological exploitation of existing network status. In this research works we have following inferences have been drawn.

The link prediction problem has been critically studied by understanding the interaction or relationships amongst the nodes across the online social network. The study particularly included questions such as how an association between the nodes was established, what are the attributes of the nodes and links actively contributes in making the existing link more sustainable, how was the impact of node properties in formulation of new links between the nodes and how the social network had been developed over the period of time.

In this work the evolution of large and complex online social networks specifically from the perspective of undirected graphs has been explored. Moreover,the existing tools and technique used for link prediction in social network were also analyzed. We have also analyzed state of the art link prediction techniques namely Jaccard Coefficient, Adamic Adar, Preferential Attachment and Resource allocation by applying them on real world online social networks such as Facebook, Deezer, GitHub and Twitch.

The exhaustive study of existing link prediction methods and models inferences that with continuously evolving online social network, the performance of techniques need to be improved. As the behavior and properties of online social network varies with its nature no technique can seldom efficient enough to address the link prediction problem in all kinds of social networks. The performance of existing techniques from complexity perspective was also affected with the scalability of networks.

In this work a probabilistic similarity link prediction algorithm is being proposed which collaboratively functions in two phases; the first phase extracts the feature of the nodes as to compute the similarity between the nodes to yield a new network with. The method computes the similarity such that it contributes to the probability of having future links between the node pair across the online social network.

In this research work we have also developed a supervised shift k-means machine learning model that classifies the potential node pairs amongst which future link can be predicted. The method exploits the local structural properties of the nodes using the classical link prediction techniques. The computed measure considered for classification of legitimate links between the nodes. The model is efficient in community detection. This proposed link prediction method supersedes other existing link prediction models in terms of complexity.

Evaluation of proposed link prediction method on the scale of performance measure: accuracy, precision and recall it is observed that the method is fairly effective and efficient in prediction of future links between the node pair across the online social networks.

Acknowledgements

First and foremost, on the successful completion of the research work and this thesis, I would like to pay all praise to the almighty God.

In this thesis the work presented would not had been possible without my close association with many people. With utmost humbleness I would like to take this opportunity to extend my deep sense of gratitude to all those who made this research thesis possible. First and foremost, I am indebted to my supervisor Dr. Aditya Khamparia, Associate Professor, School of Computer Science, Lovely Professonal University, Phagwara Punjab, for his dedicated support, timely advice, inspiration, encouragement and continuous support throughout my Ph.D. I gratefully acknowledge to my co-supervisor Dr. Babita Pandey, Assistant Professor School of Computer Science and Information Technology, Baba Ambedkar University, Amethi, under whome supervision I had started my research work.

I am grateful to my parents who always inspired me to do this work. Their affection and love can’t be expressed in words. I am thankful to my wife and life companion Mrs. Sangeeta Bhanodia for her moral support and my two daughters Pratiti and Mudra for giving me happiness throughout this work and life in general.

I am also thankful to Dr. Kamal K Sethi, Professor and Head Information Technology, Acropolis Institute of Technology and Research Indore, for his guidance in deliberating over the problems and work support.

As always it is not possible to mention everybody who had a constructive impact on this work directly and indirectly, however there are those whose support is even more important and I extent my thanks to all those peoples.

Praveen Kumar Bhanodia

Contents

List of Figure

List of Tables

Glossary

OSN	Online social network
Sociograph	Social network representation into a graph structure
Node	Element of sociograph (network graph)
Link	Connections between nodes across the network graph
Common neighbor	It is referred as the adjacent node between the node pair. Such that node v_j is a mutual node of nodes v_k and v_l such that v_j belongs to v_k and v_j both, where v_k is the set of neighbor nodes of v_k
Degree	It is the number of adjacent nodes connected to one node; for example node *u* has 4 adjacent to connected to it, the degree of *u* will be 4.
Csv	Comma separated values, it is just like an excel spreadsheet where horizontal lines are rows and vertical are columns.
Arff	Attribute relation file format, this file format is supported by Weka statistical analysis tool.
Similarity metric	The measure based on which similarity between nodes in social network is identified.
Metric	It is a calculated value exploiting attributes of graphs, for example degree of any nodes can be referred as a metric which shows about the popularity of any individual.
Path Distance	Mathematically in graph theory, the distance between nodes of a graph is the total number of edges in a shortest path, also termed as graph geodesic, or geodesic distance. If there is no connection between two nodes in the network graph the path distance would be infinity.
Homophily	People used to make friends with one who is similar to them for example people with common interest can be friends, likewise culture, age, location etc contributes in defining homophily amongst people.
Online social network	Online web application created for people to join and create community by interacting with similar peoples.
Sociogram	Social network are represented through graphs, people or users are

designated by nodes and relationship or interaction between them denoted by edges or links.

Online social network analysis It deals with the study of OSN which includes knowledge of statistic and graph theory along with understanding of physics and computer science.

Weka Open tool developed by university of Waikato for big data analysis. In our research we have used it for social network data analysis and applied machine learning techniques.

True positive It is the output which has been correctly predicted by the model to be positive. For example if the model predicts a link between the nodes and in the network it is found to be true.

True Negative It is the output where the model correctly predicts the negative class. For example if model predicts link does not exist and in the network it is found to be true.

False positive Model incorrectly predicts a link between the node pair in the social network. *i.e.* incorrect prediction of positive class.

False negative Model incorrectly predicts that link does not exist and in actual it exists. Incorrect prediction of negative class.

Accuracy The quality of the result to being precise

Precision It is also known as predictive value; it is the ratio of relevant values to the total fetched values.

Recall It is the fraction of the total amount of relevant values which are actually fetched.

Chapter 1

Introduction

1.1.Overview

The research work illustrates the thoughts proposed be Liben-Nowell and Kleinberg for addressing the link prediction problem for social networks in [1]. Link prediction is an instance of social network analysis where a link (edge) between the unconnected nodes in a sociogram or network (graph) is to be predicted on the basis of structural behavior of the graph. The prediction of link so far is approximated by exploiting the traditional social network analysis metrics to some extent. This research has explored the problem further in depth by exploiting the large online social networks namely Facebook, Twitter, Wikipedia, etc. existing around on the internet. The link prediction is probabilistic and binary classification problem in nature. Obtaining the structural features as metrics exploiting the structural properties of the network will not directly lead to predict links between the nodes across the network. In order to predict the link, the computed online social networks metrics need to be ranked orderly and then further classification approach required accordingly for a possible link between the nodes.

1.2 **Scope of the research**

The intent of this research work is to emphasize upon online social network mining referred as link prediction applied on undirected online social network represented through undirected graphs. The main aim of link prediction is to identify future missing links between nodes of the network based on previous relations that is network topology. The existing state of the art link prediction techniques also have been studied as to understand their limitations. It has been discovered that link prediction helps in addressing the issues and challenges of various domains such as recommendation of future friends. Although there are issues related to precision and scalability. In case of precision the diversity of online social networks in their nature is a challenge, because link prediction techniques used for forecasting future missing links in a specific network will not going to precisely predict the links between the nodes in another type of network. The properties of friendship network are different than co-authorship network. Moreover, precisely predicting links between nodes in sparse network is also quite difficult.

The later issue for application development using link prediction is scalability. The scale of online social network has sprawled furiously with the easy availability and last man reach of internet. Internet services are now being afforded by any other humans around us. This way the online social networks are transformed into large and complex networks generating huge amount of data. Crunching these large networks consumes most of the computational time in extracting the data and information from memory rather than processing. Therefore, the techniques developed for processing large and complex networks increased complexity levels both in terms of time and memory. In this work the problem of link prediction by exemplifying the different cases of online social networks will be discussed. Further the representation of social network data, structural feature extraction and how relevant information can be extracted to mine the network data for prediction of future links between nodes will also be discussed. Advanced machine learning and data mining techniques have been used for the purpose.

In order to explore the problem of link prediction from the perspective of approximating the potential node pair between which the link is supposed to be predicted different online social network datasets have been used. By different dataset testing it is meant that the utility of link prediction technique in general applicability across the range of online social networks.

1.3 **Main Goals**

In our work the main goals of this research work are to

1. Study the evolution of online social networks and their structures.
2. Explore the problem of link prediction.
3. Study the existing solution of link prediction in large undirected graphs, in terms of the performance.
4. Study the limitations of existing and prevailing link prediction techniques.
5. Propose, implement and evaluate link prediction methodology particularly designed for mining undirected graphs, by exploiting structural properties of the online social networks (sociography).

1.4 **Research contribution**

1. Critically analyzed the evolution of large and complex online social networks specifically undirected graphs, along with the existing tools and technique for link prediction. For this the state of the art link prediction techniques like Jaccard Coefficient, Adamic Adar, Preferential Attachment and Resource allocation have been analyzed by applying the techniques on Facebook, Deezer, GitHub and Twitch online social networks.

2. Proposed novel probabilistic similarity link prediction algorithm which extracts the feature of the nodes as to identify the similarity between the nodes. The method computes the similarity such that it contributes to the probability of having future links between the node pair across the online social network.

3. Developed a supervised shift k-means machine learning model that classifies the potential node pairs amongst which future link can be predicted. The method exploits the local structural properties of the nodes.

4. Evaluated the proposed link prediction method using state of the art performance evaluation measures namely accuracy, precision and recall.

1.5 **Motivation**

In view of internet and social network usage link prediction problem is still relatively important in the area of social network. The classic article about the problem was only being published in the year 2003 [1], since then the scale and use of online social networks has grown up in a mammoth way. And with this so does the problem of prediction. Therefore, there is scope for improvement in addressing the problem solutions pertaining to the behavior of the existing networks. The networks which were processed for link prediction were taken at a time instance and also were of small sizes that means having a smaller number of nodes altogether. So, the techniques developed at then are now unrealistic in analyzing current state of the network's structures. In previous work any significant difference between link detection and link prediction was not found, it seems both are same. Link detection and link prediction are two different problems to address and our focus is on future link prediction. Along with such gaps there are various real time applications that needs link predictions to be explored for better recommendation systems.

Link prediction has significant role in decision making; depending on the type and behavior of social networks, link prediction analysis contributes in many ways, it can be understood by following use-cases.

In the popular online social network Facebook, as known that any network is represented through sociography where in nodes are users of people and interaction between them represents through links. The type of relationship depends on the context of interaction. For instance, for a friendship edge one user initiates a friendship request which is accepted by the other node. Predicting friendship links helps in business revenue generation by posting advertisements on user pages, because knowing who is whose friend will help the business organization in identifications of user who will have a higher probability of connecting new friend across the network.

Another scenario for link prediction is in any co-authorship network [2] such as dblp, in which authors are nodes and the link represent collaboration between authors. Interaction between authors ie link between the nodes refers that there will be at least one research document published in collaboration. Analysis of such scientific networks from link prediction perspective has no direct implications, but as the network is another example of online social networks [3], an approximation of link prediction will verify the method in other networks also. Therefore, such networks also attracted the attentions of scientists, physicists and mathematicians for further study [1,4,5].

Some application examples of link prediction in real world includes following.

1. Identification of the evolution of terrorism networks and criminal network, recommendation of missing and future links in such networks. [6]
2. Recommendation of friends in a online social friendship.
3. Hypertext analysis for retrieval of relevant information across the search engines.
4. Predict the web pages which the users will likely to visit.

5. Product recommendations applications in business networks

1.6 **Problem Statement**

In online social networks the problem of link prediction is defined by understanding and representing the social network through graphs. Accordingly let us consider a social network at a time instant t where the network snapshot is designated via graph G(V,E) such that V is the set of nodes and E is the set of edges or links between nodes; such that more than one link and self-loops within the network are not allowed. Representing by universal set U containing all possible n(n-1)/2 possible links where n is the number of nodes (node set), here set of node is designated by V. Then, the set of nonexistent links is $U - E$. It is assumed that there will be certain future missing links (or links that are likely to appear in the near future) in the set $(U - E)$, and approximation of such missing links is termed as link prediction.

1.7 Research Objectives

The specific objective drawn for the research work is as follows.

Objective 1: To analyze the existing approaches available for link prediction in social network.

Objective 2: To propose a novel method for link prediction in social network.

Objective 3: To implement and evaluate the proposed approach.

1.8 **Research Questions.**

Evolution and development of online social network is fast with respect to time and it is quite challenging to understand itself evolution and the link prediction problem is one instance of it. Addressing link prediction problem by analyzing online social networks will further help us in development of better solution for various intelligent expert systems that used to recommend desirable and effective results. The questions used to seek researchers attentions are like:

1) How an online social network evolved?
2) What is an online social network and how it can be represented?
3) What is social network analysis?
4) What is link prediction in online social networks?

5) What are the different techniques used for prediction of potential link between nodes, how machine learning could contribute in predicting links between nodes in online social network?
6) How link prediction could be understood and addressed?
7) What is the pattern of relationship between the nodes and how does it change with the scaling of the network?
8) What is the impact of other nodes over the relationship of two nodes?
9) How does the relationships and interaction between nodes will impact to form new links between the nodes?

In this research work prediction of future links between the node pairs in an online social network has been addressed. The problem refers to predict the future possibility of establishing an association between node pairs, such that there is no association amongst the nodes in the present state of network. In this research work we have proposed a link prediction technique which measures the similarity between two potential nodes by exploiting the local node neighborhood property. The measure thus computed contributes to the probability of predicting a link in near future. In this research work a machine learning based model, based on shift k means algorithm is also developed for converging the large complex online social network to the cluster of highly ranked node pairs between which there will be a strong possibility of link existence.

1.9 Social Networks

In 1954, Barnes had used ‘Social Network’ at first to represent the human relations in Committees in Parish Norway [7,8]. Social network is a structural network containing nodes representing individuals, organization or an entity and the link connecting the nodes represents certain social relationships between the individuals. Figure 1.1 demonstrates a sample social network of nodes and connections.

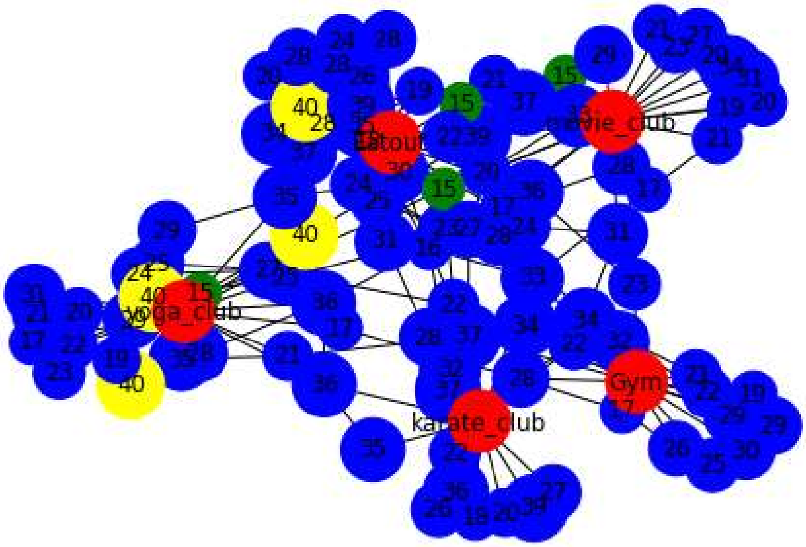


Figure 1.1. A Social Network Example

Each node in the network is a unique entity which can be a person, place or group. Links which are actually connections between these nodes are relationships between them; these represent relations with family members, professional friends and so on. Thus it is at large is a kind of social structure consist of nodes and links displaying interdependencies on each other [9]. A long back ago social networks were evolved by having physical connections amongst people, now it is evolved virtually through several social networking applications. Development of web technology with availability of internet has eliminated the demographic limitations or constraints of old age traditional social networks. There are many popular online social networking websites prevailing on the internet such as Facebook, LinkedIn, Twitter, Hi-Fi, etc.[10]. Online social networks are dynamic in nature, because addition and deletion of new links and nodes are used to happen with respect to time. In addition to this the according to the nature of the network, they are often diverse in nature, therefore analysis of a particular network will give information about the network progression itself. Structural exploitation of the network will play a significant role in analysis of these online social networks. Link prediction techniques are instrumental in such analysis across several application domains [11-16].Prediction of links between the nodes involves measurement of similarity between the nodes across the entire network [17-21]. Computation techniques of such similarity measures for link prediction are referred as link prediction techniques.

Techniques used for link prediction may be used for many applications across domains like possibility of having some mutually connections between academic, industrial or research professionals in field of research, academics and industry [22], the historical navigation data analysis will help in generating tools that are more efficient in further navigation recommendations [23], it will overcome the data sparsity problem in various recommendation systems, etc.

1.10 Thesis organization

The thesis is organized as under:

Chapter 2 Link prediction background

The chapter 2 illustrated the previous literatures available on prediction of links in social networks. The research methods, algorithms and models proposed for prediction of possible links between the nodes pairs in different types of online social networks have been discussed exhaustively. The chapter also describes the existing link prediction techniques with appropriate examples and critical evaluation.

Chapter 3 Methods and Materials

Chapter 3 discusses the methodology employed for link prediction between node pairs in large online social networks. The algorithmic research methodology adopted has been discussed in detail. The chapter describes the probabilistic link prediction technique used to determine similarity between the nodes to approximate the probability for having a link between them. Further it also describes a supervised shift k means machine learning based model for link prediction which is used to converge to potential nodes between which link is predicted.

Chapter 4 Experimental Analysis

Illustrates the experimental setup along with analysis required to implement the proposed algorithms and model for link prediction. It describes the real online dataset downloaded for testing and validation of the link prediction techniques. It also describes the tools and technology used for implementation and analysis of online social network data for further link prediction. The chapter describes the results and outcomes generated when the techniques is applied on the real online social network dataset of Facebook, Twitch, Wikipedia, Github and

Deezer. The performance metrics obtained has been critically analyzed and the technique proposed has been evaluated using classical performance evaluation parameters.

Chapter 5 Conclusion

The chapter 5 concludes the research work undertaken for link prediction problem in online social networks. This chapter discusses the key outcomes for the objectives of the research work along with future challenges and research opportunities.

1.11 Summary

In this chapter our objective was to introduce about the links prediction problem in online social networks where in prediction of future link between the nodes is the sole objective of the research undertaken. The chapter encompasses about the research contributions and motivation for addressing the issues and challenges of link prediction in social networks. The chapter also systematically defines the link prediction problem statement. It also briefly introduces about certain questions related to investigation of this topic.

Chapter 2

Basic Concepts and Link Prediction Background

2.1 Overview

In the first chapter, a brief introduction about the link prediction problem in online social networks is given. It has been understood that, what social network analysis is and how link prediction is an instance of this analysis. Also explored the various applications of link prediction in solving variety of social and professional problems, various recommendation systems around us are one of such real time applications. This chapter provides background of link prediction in online social networks, particularly its definition and the methods proposed to address the problem. In order to understand the problem effectively relevant graphs have been created. The chapter also explores the evolution of social network along with its representation through nodes, links between the nodes. Moreover the techniques and metrics used for prediction of links and corresponding relationship between the node pairs across the network are thoroughly reviewed. It is worth mentioning here that we will discuss the classical link prediction techniques in detail and will not discuss all of the existing methods as they are not in our scope of research.

2.2. Online Social Networks (OSN)

People tend to have relations with each other and similar people are more likely to have connection in between and form a group, such groups are referred as social networks. Since the inception of internet and reach of World Wide Web applications at use for people, websites meant for such social connections are known as online social networks (OSN). The connection between two people refers to various kinds of relationships between them. The mathematical modeling of online social network as to understand the structure consists of nodes and links known as graph theory [24][25].

2.3. Types of online social network

The online social networks broadly categorized in two ways: 1) certain network or 2) uncertain networks. In uncertain networks, the probabilities of link prediction are associated with the interaction between the nodes and in certain type there are no probabilities associated with the interaction that is link between the nodes. These certain and uncertain online social networks are further bifurcated as: static and dynamic [26][27] and dynamic is further classified as: incremental, decremental or mixed.

2.3.1 Static Network

On the idea of study, it's been observed that static network is often categorized in two groups: certain network and unsure network. The networks are represented by connected nodes through edges. The node of static network never crashes down nor changes its position. The sides or links maintains the operational status all the time. The whole structure of the network will remain same. for instance , an instance of a social network on specific time are going to be a sort of static network. Unlike certain static network in uncertain static network nodes are connected by edges with some probability.

2.3.2 Dynamic Networks

The Dynamic online social networks used to alter their structure with respect to time. These networks are further classified according to three categories as demonstrated in Figure 1:

1. Change in number of nodes: nodes across the network may visible and invisible with respect to the time (also known as stochastic network)
2. Total number of Nodes is fixed only edges get crashed and recover accordingly.
3. The total number of nodes in the network is fixed but over the period of time the networks is evolved with new links formed between the nodes and accordingly positions of the nodes has been changed sometime.

Further online social networks of static category can further be classified as:

a) Incremental social network
b) Decremental social network

c) Hybrid social network

Incremental social networks are such networks in which new links between the nodes are appeared leading to an increase in the nodes and links, therefore it is incremental in nature. In the contrary in decremental over the period of time existing links will be disappeared due to certain noise or inactive interaction between the nodes. In hybrid the process of link creation and elimination happens simultaneously or we can say it is the combination of both the former type of network [28].

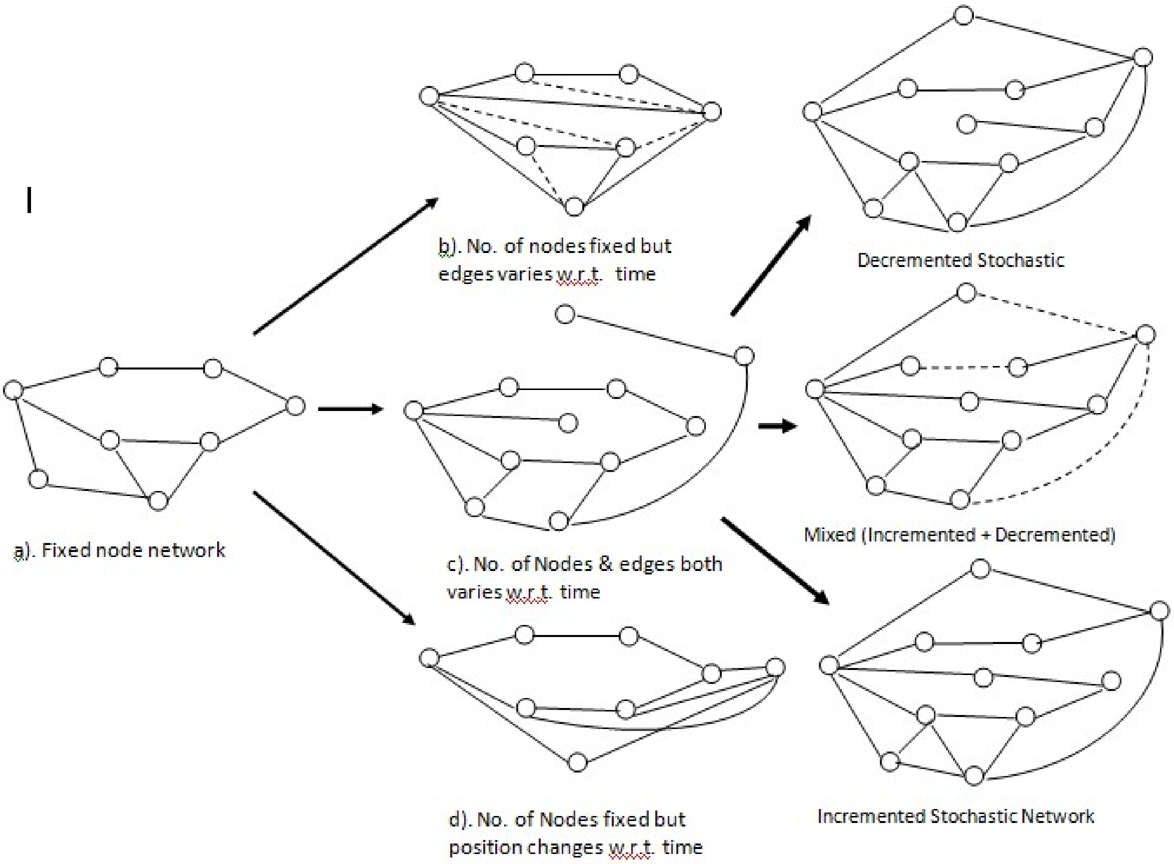

b). No. of nodes fixed but
edges varies w.r.t. time
Decremented Stochastic
a). Fixed node network
c). No. of Nodes & edges both
varies w.r.t. time
Mixed (Incremented + Decremented)
d). No. of Nodes fixed but
position changes w.r.t. time
Incremented Stochastic Network

Figure 2.1:- Different types of network representations; a) Static network; b) describes evolution of network, here nodes are fixed but links will vary with time; c) describes evolution of network, where number of nodes and edges both varies with respect to time; d) describes evolution of network, here the number of nodes are fixed but their position changes with time, e) Decremental stochastic network: few nodes eliminated, f) Mixed network: few existing nodes deleted and new will be added, h) Incremental network: new nodes and links are added.

2.4. Link Prediction Problem

As demonstrated in Figure 2.2, the link prediction problem can be understood in way that over the period of time online social networks used to grow as the numbers of new users (nodes) are addressed and subsequently new interactions between nodes across the network also get developed. Therefore accordingly link prediction with respect to time the edges or links between nodes is to be approximated on the basis of the pro data or information available.

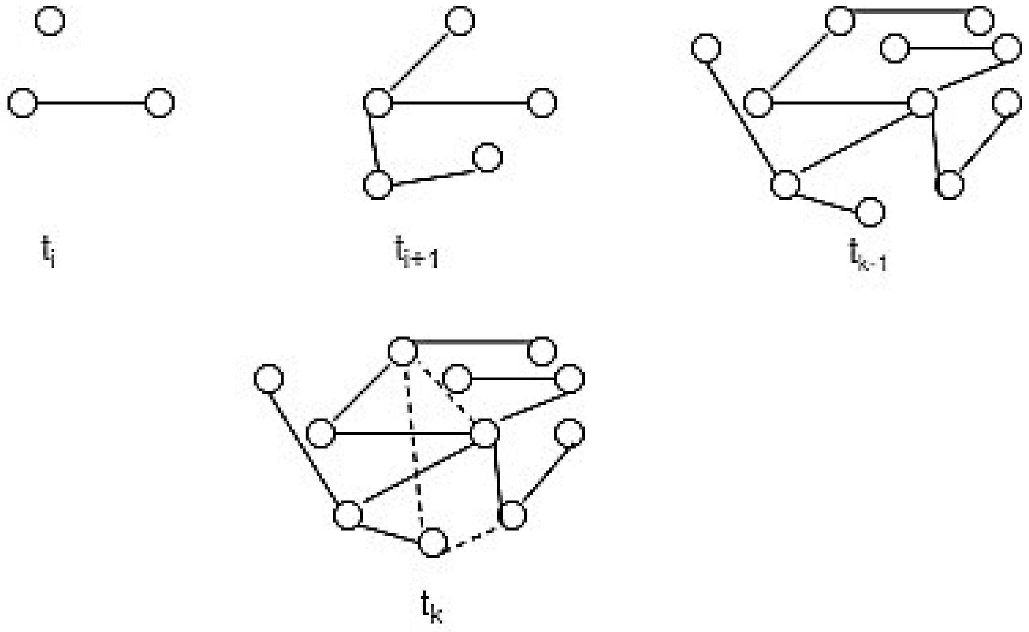


Figure2.2. Link prediction problem illustration

2.4.1. Link prediction problem illustrations

As illustrated in [29] the typical link prediction problem depending on its existence and dynamism of the social networks is illustrated mathematically as under:

A set of data at an instance is given as *G=(E,X)*; E is the set of observed links such that *E={E$_i$}* where i varies from 1 to n. Whereas X is the set of unobserved links which is not existing between the node pairs, such that x_{ij} does not belongs to E. Therefore, link prediction is the process to compute the probability of existence of the unobserved links on the basis of network properties.

Based on the type of dynamism of online social networks and the methods used to predict the link, the link prediction problem is formulated in different ways. A classic definition of link prediction problem is expressed by the snapshot of network at any given time t. Consider an undirected network $G\ (V, E)$, where V is the set of nodes and E is the set of links. Multiple links and self-connections are not allowed. Denote by U, the universal set containing all $|V| \cdot (|V| - 1)/ 2$ possible links, where $|V|$ denotes the number of elements in set V. Then, the set of nonexistent links is $U - E$. Assume that there are some missing links (or the links that will appear in the future) in the set $U - E$, and the task of link prediction is to find out these links.

In case of stochastic increment network, In a given graph $G(V, E)$ at time t, If at any time $t' > t$, set of V'_t nodes and E'_t links are appeared in the graph G , then the new incremental graph at time t' is $G_{t'}\ (V_{t'}, E_{t'})$ where $V_{t'} = V \cup V_{'t}$ and $Et' = E \cup E'_t$.

In case of the contemporary stochastic decrement network, given graph $G(V, E)$ at time t, If at any time $t' > t$, set of $V''t$ nodes and $E''t$ edges are disappeared from the graph G, then the new decremental graph at time t' is $G_{t'}\ (V_{t'}, E_{t'})$ where $V_{t'} = V \cap V'_t$ and $E_{t'} = E \cap E_{'t}$.

In case of stochastic mixed network, For the given graph $G(V, E)$ at time t, If at any time $t' > t$, set of V'_t nodes and E'_t edges are appeared in the graph G and V''_t nodes and E''_t edges are disappeared from the graph G , then the new mixed graph at time t' is $G_{t'}\ (V_{t'}, E_{t'})$ where $V_{t'} = (V \cup V't) \cup (V \cap V'_t)$ and $E_{t'} = (E \cup E't) \cup (E \cap E''t)$.

For this graph $G_{t'}$, U denotes the universal set containing all $|\ Vt'| \cdot (|\ Vt'| - 1)/ 2$ possible links, where |V| cardinality of set $V_{t'}$. Then, the set of nonexistent links is $U - E_{t'}$. Assume that there are some missing links (or the links that will appear in the future) in the set $U - E_{t'}$, and the task of link prediction is to find out these links.

Uncertain network G at time t is denoted as four tuples $G(V, E, P,A)$, where V *and* E are same as in the classical definition of graph. P is probability associated with each edge belongs to graph G *and* A is adjacency matrix of NxN, where $|V|=N$. At time t', such that $t<t'$ the link prediction problem in uncertain graph is to predict the probability of occurrence of U-E edges in graph.

The time series link prediction problem is defined as: Let N be the list of nodes, $N = \{1, 2\ldots, n\}$. A graph series is a list of graphs $\{GH_1, GH_2,\ldots,GH_t\}$ corresponding to a list of adjacency matrices $(MT_1,MT_2,\ldots,MT_t)$. Each MT_t is a $n \times n$ matrix with elements $MT_{t(i,j)}$ corresponding to the edges in $ETt_{(i,j)}$.The value of $MT_{t(i,j)}$ is from the set $\{0,1\}$, o means link (i,j) does not exist and 1 indicates link (i,j) exists during the period t. Then in the time series link prediction, predicts the existence or non-existence of the links in time $T +1$ using previous times $MT_1,MT_2,\ldots,MT_t$.

2.5 Link Prediction Taxonomy

The methods used for link prediction are majorly classified on the basis of four parameters which as under:

1. The technique used for link prediction
2. Features used for link prediction
3. The network on which the technique is supposed to apply.
4. How the problem is formulated.

The prediction techniques are categorized as artificial intelligence based or similarity based. Either of the technique depends to work upon local similarity index, proximity index, social features, temporal features or hybrid features for prediction of the link, the only difference between these two is that (artificial intelligence) AI based techniques gives abstract view and needs less computation rather the later needs more information and more computation facility. Link prediction is also categorized as inter link which means prediction of links between the node pair exists within the community whereas the other one is intra link which means across the communities. The features extracted using node attributes as local information is mostly used in the case of dense network. Local similarity index is being used for approximation of link prediction most of the time. The detail taxonomy and dependency of link prediction problem in online social network has been demonstrated in Figure 2.3.

The thorough study of literature on online social network, online social network analysis and the link prediction problem in online social network has inference that the link prediction problem has been addressed by several researchers and accordingly the techniques have been categorized in the figure. However the following section illustrated the link prediction in detail.

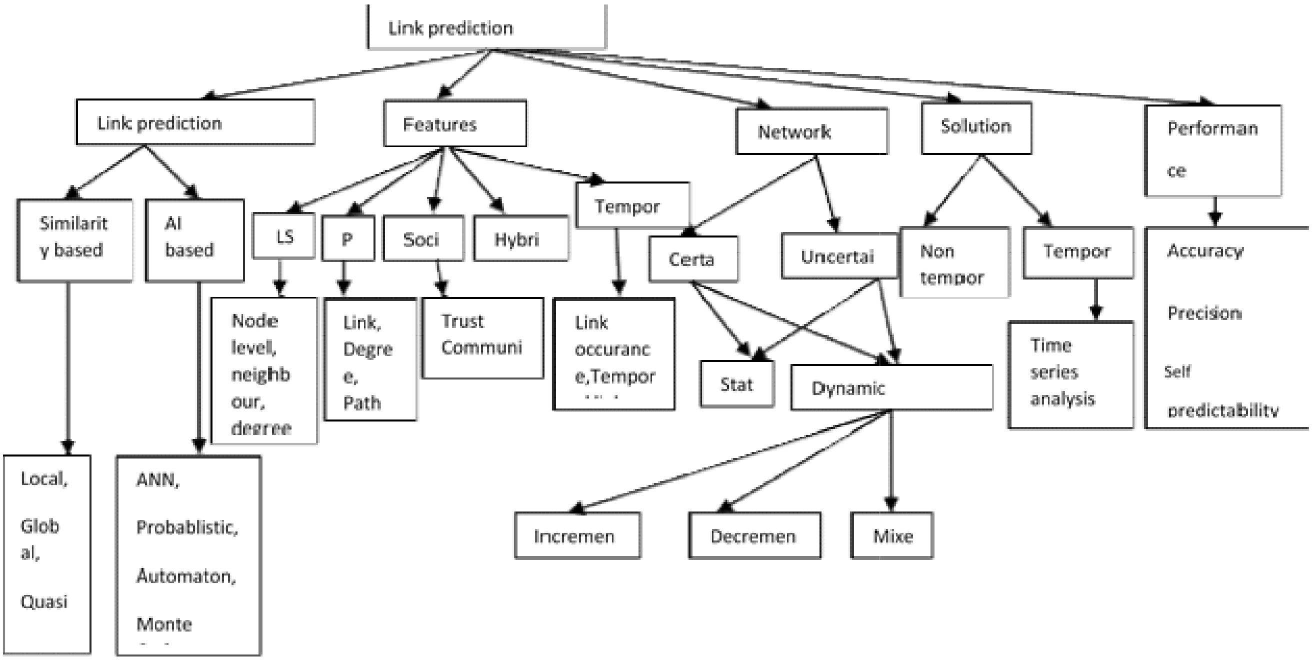


Figure 2.3: Detail taxonomy of link prediction problem categorized according features, prediction methods, network types, solution proposed and performance of the techniques.

2.6 The Link Prediction Techniques

We have classified the similarity based link prediction methods based on the study as: local link prediction techniques, global link prediction techniques and Quasi methods for link prediction.

2.6.1 Local Similarity Methods

The Local similarity-based methods generally use the local structural information of the online social network components in order to calculate the similarity between the node pairs as to predict the future link between them. These local similarity measures are also known as local similarity metrics or index. The node neighborhood property is being extracted for calculating the measure. The obtained value is then observed if it is more than the threshold value the probability of having a link between the nodes will be more. The local link prediction techniques are as follows.

2.6.1.1 Common Neighbors (CN): The Common neighborhood method gives a measure based on which the similarity between the nodes is calculated. It refers to the intersection of the sets of neighbor nodes to the node pairs for predicting future link is computed. For example, in Figure 2.4 the nodes are the social network components used to refer the users, items, etc. in any online social network. Similarly, the link or edges between these nodes represents the interaction or association or relationship between users in a network. The dotted line represents future relationship/link which is to be predicted while solid lines are existing relationships [17]. The mathematical computation of common Neighbors (CN) is calculated as follows (refer equation 2.1).

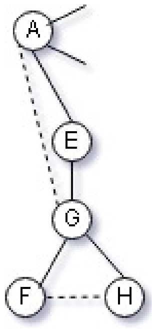


Figure 2.4: five node graph for CN and JC demonstration

$$Common\_Neighborhood_{(x,y)} = \Gamma(x) \cap \Gamma(y) \text{------------------------}(2.1)$$

It is to be understood that $\Gamma(x)$ and $\Gamma(y)$ designates mutual neighbors nodes between node x and y.

In figure2.4 the common neighborhood similarity for node A and G is CN (AG) is 1 and similarly the similarity measure CN for node F and H is 1. Here in this case it quite difficult to decide the probability of having and interaction between node pair AG or node pair FH as similarity measure for both the node pair is same. The weighted Common Neighbors (CN_w) is computed as follows where $w_{(x,y)}$ is the number of interactions between the nodes x and y, it can be computed using equation 2.2.

$$CN_{w(x,y)} = \sum_{z\epsilon(x)\cap\Gamma(y)}^{\infty} \left(\frac{\mathrm{w(x,z)\ +\ w(y,z)}}{2} \right) ------------ (2.2)$$

2.6.1.2 Jaccards coefficient (JC): Jaccard Coefficient for link prediction in social network used to calculate number of the common neighbors between the node pairs where future link is supposed to be predicted. Let us consider Figure 2.4 for understanding how this similarity metric works. In CN there are cases where the intersection between the adjacent numbers of nodes of the node pairs could be same. In such cased it will be difficult to approximate and predict the existence of future link using the technique, and thus JC is the similarity measure that could produce a normalized score [1]. Mathematically it is computed according to equation 2.3 given as under.

$$JC_{xy} = \frac{\Gamma(x) \cap \Gamma(\mathrm{y})}{\Gamma(\mathrm{x}) \cup \Gamma(\mathrm{y})} ---------------- (2.3)$$

2.6.1.3 Adamic-Adar (AA): According to AA similarity metric [18], similarity index for node pairs between which links to be predicted is calculated by exploiting the common attributes of the node pairs. For example in Figure 2.5, the dashed lines are the links to be predicted and smooth lines are the existing links. Accordingly the index for node pair AC and AE is calculated as; AA (AC) = 1/log2 + 1/log2 and AA (AE) = 1/log2 +1/log2, the obtained value contributes in the prediction of link between the node pairs. Mathematically it is computed as follows (Refer Equation no. 2.4).

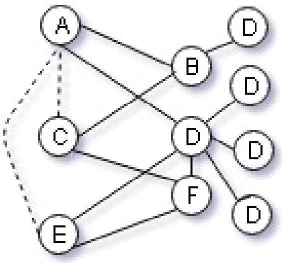


Figure 2.5 : Adamic/Adar Demonstration

$$AA_{xy} = \sum_{z\epsilon(x)\cap\Gamma(y)}^{\infty} \frac{1}{\log\ |\Gamma(z)|} ------------------ (2.4)$$

2.6.1.3 Resource Allocation (RA): The resource allocation link prediction measure based on the distribution of resources across the adjacent connected nodes. It is extension of Adamic-Adar similarity measure [30]. According to the method the quantum of penalty (which is *1 /logk(z)* and *1/ k(z)*) to the higher degree node differentiate it from AA. In addition, it is to be understood that the difference between the two is insignificant for the nodes where the average degree is reasonable less. It is computed as 1 divided by sum the mutual neighbor nodes of node pair, for large mutual neighbors the resource allocation measure would also be high. The mathematical definition of resource allocation link prediction measure is shown in equation 2.5.

$$RA_{xy} = \sum_{z\epsilon(x)\cap\Gamma(y)}^{\infty} \frac{1}{\mathrm{k(Z)}} ---------------(2.5)$$

2.6.1.4 Salton index: Salton similarity measure was introduced by Salton [31]. It is developed to computer the ration of intersection of degree of the two node pairs to root of product of number of adjacent nodes of the node pairs. Mathematically it is defined using equation 2.6

$$S_{xy} = \frac{\Gamma(x) \cap \Gamma(\mathrm{y})}{\sqrt{}\mathrm{k(x)} \times \mathrm{k(y)}} ------------------ (2.6)$$

It should be understood that *k(x)* =*|Γ (x)* | is the degree of node *x*. this similarity measure index is also known as cosine similarity.

2.6.1.5 Preferential Attachment (PA): According to Newman [31], preferential attachment identifies the similarity between two nodes by computing the product of their degrees, higher the product value, higher will be the probability of having a link between them. For example refer figure 6, the degree of node G is 3 and that of node F is 7, hence the chances of future link between node A and node F is high then node pair AG. Mathematically it is computed as follows (Refer equation no. 2.7)

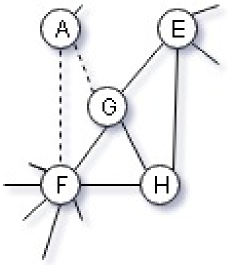


Figure 2.6: Preferential Attachment Demonstration

$$\mathrm{PA}_{(x,y)} = \Gamma(x).\Gamma(y) - - - - - - - - - - - - - - - - (2.7)$$

The calculation of preferential attachment measure for AF = 1* 7 = 7 and for AG = 1 * 3 = 3, it is to be noted that PA(AF) > PA(AG).

2.6.1.6 Sørensen index: This similarity measure was basically developed for ecological community data [32]. Equation 2.8 represents it mathematically which is as follows:

$$S_{xy} = \frac{2 * |\Gamma(x) \cap \Gamma(\mathrm{y})|}{\mathrm{k(x)} + \mathrm{k(y)}} - - - - - - - - - - - - - - - - - - (2.8)$$

2.6.1.7 Hub Promoted Index (HPI): This link prediction measure used to quantify the structural overlapping of node pair substrates specifically in metabolic networks [33]. It is calculated as twice of the intersection of the degrees of the node pair to the minimum number of adjacent nodes connected to either one of the two nodes. Mathematically it is calculated as (Refer equation 2.9):

$$Similarity_{xy} = \frac{2 * |\Gamma(x) \cap \Gamma(\mathrm{y})|}{\min|\mathrm{k(x)}, \mathrm{k(y)}|} - - - - - - - - - - - - - - - (2.9)$$

2.6.1.8 Leicht–Holme–Newman Index (LHN): This similarity is similar to HPI similarity index the difference is it considers the product of degree of the nodes of the nodes between which link is to be predicted[34]. Mathematically the ratio is computed as under (Equation no. 2.10):

$$Similatity_{xy} = \frac{2 * |\Gamma(x) \cap \Gamma(\mathrm{y})|}{\mathrm{k(x)} * \mathrm{k(y)}|} ---------------(2.10)$$

2.6.2 Global Similarity Methods

Previous section illustrates the link prediction methods that exploits the local topological properties or attributes of the online social network represented through graphs. Certain methods have been developed and proposed by researchers that consider the other attributes as well are referred as global similarity methods. These method extract all the relevant information from the social network structure for calculating the similarity values between node pairs.

2.6.2.1 Path Distance: According to the method it considers the path between the two nodes where link is to be predicted. This distance in network is an obvious metric for identification of the proximity of node pairs, in literatures sometime also referred as geodesic distance. Using Dijkstra's for fetching the shortest path between the node pairs will not be quite effective in online social networks. Although it is efficient in small networks. It can be extended by using ring search property for computing the shortest path or distance for the node pair between which link is not existing. A value thus computed helps in analyzing for approximation of the prediction of the future links between the node pairs.

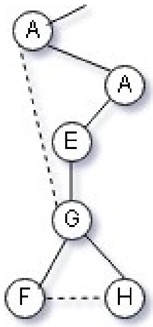


Figure 2.7: Path Distance

For example consider Figure 2.7 where the path calculated between node pair AG = -3 and node pair FH = -2, and it is observed that FH > AG, hence a link between node pair FH is more likely to occur in future. It is to be understood that a negative signed value for the shortest path designates that proximity between the node pair increases with the closeness of the node *x* and node *y*.

2.6.2.2 Katz (Exponentially damp path counts): It is calculated by taking into account all the paths available between the node pair and rate such short paths more strongly. As to assign less weight to loner paths the computation decreases the penalties to the entire measurement. For this it uses β_l as a factor such that l is the length of the path. The mathematical definition can be referred in equation no 2.11.

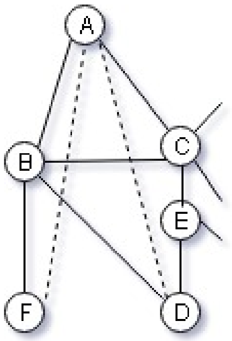


Figure 2.8: Katz Demonstration

$$Katz(xy) = \sum_{l=0}^{n} \beta < path(x,y)\text{---} \textit{(2.11)}$$

The β value controls the length, small β contributes less as length of three or more considered less into account and thus the method converge the mutual neighborhood technique. The complexity is roughly cubic because it needs a matrix inversion [17][35]. For example in Figure 2.8 the damp path count for $AD_1 = 3$ and $AD_2 = 3$; path count for $AF_1 = 2$, path count for $AF_2 = 3$ and path $AF_3 = 5$. Therefore AD = 3+3 =6 and AF 2+3+5 = 10.

2.6.2.3 Measurement index in weighted networks

Weighted social networks shows some extra features, therefore its mathematical representation has been modified accordingly (Refer Equation no. 2.12, 2.13 and 2.14), thus computation of similarity measure for the node pair between which links is supposed to be predicted and henceforth the methods for doing so are as follows:

Weighted common neighbor (WCN)

$$\boldsymbol{S_{xy}^{WCN}} = \sum_{z\in O_{xy}} \boldsymbol{w(x,y)^{\alpha} + w(z,y)^{\alpha}} - - - - - - - - - - - - - - - - - (\mathbf{2.12})$$

Weighted Adamic/Adar (WAA)

$$\boldsymbol{S_{xy}^{WAA}} = \sum_{z\in O_{xy}} \frac{\boldsymbol{(w(x,y)^{\alpha} + w(z,y)^{\alpha})}}{\boldsymbol{log(1+s(z))}} - - - - - - - - - - - - - - - (\mathbf{2.13})$$

Weighted Resource Allocation (WRA)

$$\boldsymbol{S_{xy}^{WRA}} = \sum_{z\in O_{xy}} \frac{\boldsymbol{(w(x,y)^{\alpha} + w(z,y)^{\alpha})}}{\boldsymbol{s(z)}} - - - - - - - - - - - - - - - - - - - (\mathbf{2.14})$$

Where O_{xy} is the set of mutual neighbors nodes between node *x and y*; $w_{(x,z)}$ designates weight of link between the node pair; $\boldsymbol{S(x)} = \sum_{z\in\Gamma(x)} \boldsymbol{w(x,y)^{\alpha}}$ **x** and z,. In addition for α = 0, *s(x)*= degree of node *x*, calculated for unweighted indices; for *α=1*, the calculation for weighted indices. In General, optimal α is less than 1 for most of the weighted networks.

2.6.3 Quasi-Local Matrix Methods

These link prediction methods does not consider global topological information for identification of closeness between the node pairs for approximating a future link rather exploits more features and information from local.

2.6.3.1 SimRank: According to SimRank link prediction technique it ranks the degree of similarity between the two nodes. The basic idea is that two nodes will be similar if they are related to similar kind of objects. It can be understood that two nodes are similar if have same neighbor node. For example node a and node b are similar if both are connected to x and y assuming that x and y both are similar to each other [36,37]. It is measured on the scale of o and 1; 1 means maximum similarity and 0 denotes completely in similar. Node not connected to any of the other nodes in the social network will have similarity 0.

2.6.3.2 Hitting Time and Commute Time: A random walk on a given graph moves iteratively over the graph from a node x while selecting each path at random for the next node. The expected number of steps through a random walk to get from x to y is known as the Hitting Time H(x, y). A short hitting time means similarity of the node and therefore a higher probability of future connecting. The commute time C(x, y) is a hitting time variant i.e. useful for undirected graphs since the time of hitting is not symmetrical. Mathematically hitting time and commute time is calculated as follows (Refer Equation no. 2.15).

$$C(x,y) = H(x,y) + H(y,x) \text{---} (2.15)$$

It is observed that the commute time may be high in terms of variance therefore link prediction using may get affected using this feature. Let us suppose that a node z is having high stationary probability of node x and y, a random walker then perhaps reach to the adjacent node of z. in order to avoid this walker can be reset to x with fixed probability of α.

2.6.3.3 Rooted Page Rank

A single vertex attribute measurement of a page rank is extended to the rooted page rank to predict a relation. In order to return *x*, it is determined as the quantity of measures from *x to y* with α probability. By adding the counterpart where the position of *x* and *y* is flipped, the rooted page rank is asymmetric and translated into symmetric [38]. Mathematically it can be calculated according to equation (2.16).

$$D(i,j) = \sum_j A(i,j) \text{ ---} (2.16)$$

2.6.3.4 PropFlow and High-Performance Link Prediction

According to this the similarity measure between two nodes is identified by approximating the success probability of random walk starting from source node to the destination node would be not exceeds than steps l. The technique is more local as compared to rooted page rank prediction technique although this is inspired form page rank algorithm. The techniques are faster as it does not need random resets.

High-performance Link Prediction as a framework for link prediction and distinguished in two categories.

- HPLP: It is not supposed to exploits the prevailing unsupervised technique; it is only a simple measure such as Indegree and Out-Degree, Max. Flow, Shortest Paths or PropFlow
- HPLP+: It uses full feature set adding Adamic/Adar, Jaccards coefficient, Katz and Preferential Attachment [39].

2.6.3.5 Supervised Random Walks

For link prediction, topological information with node and link features will be used, and the related information and attributes will be used for future prediction of new links. The edge strengths available here are used for further learning of the system to assess the strengths or probability for possible links between the nodes.

2.6.4 Node attributes based methods

There are techniques and methods that exploits attributes of nodes (NA), such methods use node or link attributes for predicting future links like: collaborative filtering [40,41], stochastic relational model [42] and iterative collective classification[43].

2.6.4.1 Collaborative Filtering

Two or more different link prediction techniques or methods are implemented in collaborative filtering on online social network dataset where the links will be ranked (may be 5-10). In addition to gathering item rankings, it also takes advantage of other features (genre reviews in the case of movies).collaborative filtering techniques are user-based and item based; user-based filters are defined as the connection or closeness between users. Once the similarity between users is predicted and suggested accordingly, item based similarity is predicted on the basis of metadata of objects and related products of interest are then recommended to the users.

2.6.4.2 Stochastic Relational Model

Stochastic relational models are connection wise processes generated by the interplay of tensors of much entity wise Gaussian processes. Basically these models are defined as a set or group of non parametric priors on dimensional tensor matrix of infinite nature, each element of the set denotes the relation amongst entities tuples. By optimizing marginalized likeliness, the information will be exchanged between Gaussian processes via the whole relational networks, such that link dependency structure will be shared to the entities, represented through the learned Gaussian process kernels [44].

2.6.4.3 Iterative Collective Classification

The node attributes and edges among them benefit from collective node classification and edge prediction. If, for some cause, the output of the method is adversely affected in complex social networks if nodes are missing at an instant. Iteratively, both link prediction and the classification of collective nodes performed in order to improve. The classification of nodes uses the exiting node and link related information and provides inferred link prediction information. Because of this, the link prediction process that exploits exaggerated previous attributes which are unknown will be able to boost its performance. Similarly, link prediction also exploits the existing node and link knowledge, as to predict the earlier links for unknown nodes. In addition the new links will be added in cases where the newly approximated link modifies the topological attributes used for further classification of the legitimate links between the nodes. Thus the output of the collective node classification improves.

2.6.5 Correlation information-based methods

This technique uses node to node, link to link or node to link correlation information for prediction of links between the nodes in a sociograph. Based on this there are certain link prediction techniques proposed by researcher like cold-start approach [45], edge coefficient generation [46] and indirect global silencing of correlations [47].

2.6.5.1 Cold-start edge prediction in multi-relational networks

The cold start approach for link prediction is based on a latent or hidden space network model. The latent space is quantified by fetching a low-dimensional factor the adjacency matrix of *NxN* dimension. It uses probability ratio tests for determining the association among the sub-networks via latent or hidden factors. Apart from this, a regression is generated on the associated variables. By approximating the significance of the node, the method will be focused on the association of existing nodes of the network to the new incoming nodes.

2.6.5.2 Global silencing of indirect correlations

The link predictions between the nodes in social networks are usually based upon association of experimental calibration like gene expression, especially for complex biomedical cellular networks, and are apparently influenced by direct and indirect paths. For building a technique to suppress the indirect results, it utilizes this fundamental property of dynamic correlation. For changing the correlation matrix to discriminative, it uses matrix transformation. This reinforces the terminology associated with direct causal connections. In the translation of the correlation results, this silencing approach will help to explain the interaction insights of the system. The method is therefore extended to complex structures that explicitly control biological networks over edge prediction.

2.6.6 Most influential node identification-based methods

These approaches are based on various approaches to classify the most prominent node, such as: de-anonymization [48], learning spectral graph transformations [49], ranking factor graph model and transfer-based [50], ORFP (without and with game theory) [51] and balanced modularity maximization model [52].

2.6.6.1 De-anonymization

The fundamental de-anonymization method is defined [48]. The first seed identification and the second propagation are followed by two distinct phases. The first step involves de-anonymizing small numbers of nodes, and these nodes are used as anchors in the latter step to spread the de-anonymization to more and more nodes.

2.6.6.2 Ranking factor graph model (RFG) and transfer based RFG

The technique inspired from intuition of individual people befriending with individuals possessing similar kind of values in different networks. This is focused on the understanding of many social phenomena that are typically prevalent across heterogeneous networks. A transfer-based RFG model that incorporates knowledge about the network structure is built with these general patterns. The model offers insights about the basic concepts used to guide the network's link creation and evolution.

2.6.6.3 Diverse node adoption algorithm

In order to understand the generation of a link between two nodes, this approach used the nodes' evolution diversity factor. The diversity of node evolution is based on the premise that different nodes in a social network which have different mechanisms or processes to produce different link types.

2.6.7 Artificial Intelligence Methods

The use of AI method in Social Network has been documented by many literatures. All these literatures stated that either for network representation, classification or prediction, the AI method in Social Network is deployed [53]. A significant subject of research is social network representation. Social Networks are represented through matrix or graph. Network size influences the efficiency of methods for link prediction. Artificial intelligence-based approaches are implemented for social network representation, such as: automata [54] and Deep Belief Network [55]. It uses depth-based representations to reduce the complexity regarding the computation of the graph. For weight adjustment of links or edges, Dempster-Shafer theory [56], Deep Learning [57] and Artificial Neural Network [58] had been used, and in the same fashion for learning and prediction, Markov chains and automata are used. These methods based on artificial intelligence subsequently minimize the computational overhead, complexity and cost of the process of link prediction.

2.6.8 Hybrid Methods

There are methods which are combination of various techniques for prediction of future links between the nodes in a online social network. The approach may contain combination of two or more favorable link prediction methods.

2.6.8.1 Evidential Measurement

Evidential measurement was proposed by Yin et al., for determination of similarity between the node pairs the measure needs (Yin et al., 2017)[58] both local similarity technique and node similarity techniques. The mathematical computation to obtain the measure is as follows. Refer equation 2.17 given below.

$$\mathbf{S_{i,j}} = \sum_{z\epsilon(x)\cap\Gamma(y)}^{\infty} \frac{\boldsymbol{\varphi_{i,j}}}{\boldsymbol{\phi_{z}}} \text{------------------------} (2.17)$$

It is to be understood that, φ_{ij}designates the structural similarity and ϕ_{Z}designates the similarity computed by attributes.

2.6.8.2 Collaborative Filtering link prediction approach

Social network relation prediction is tackled through mutual filtering, these two different recommendation approaches are applied to user data where the objects are rated on a scale (may be on a scale of 5 to 10). Generally, a movie recommendation scheme is based on such collaborative filtering. In addition to gathering item rankings, it also takes advantage of other features (genre reviews in the case of movies). In a user to user friendship, similarity defined through the two combinations of filters: item based filter and user based filter. The identification of similarity on the basis of user is done, accordingly friendship will be recommended to an individual. In case of item based similarity, it is identified on the basis of meta data of the objects and prediction of related products is performed accordingly.

2.7 Performance Evaluation Metrics

The link prediction methods and models are evaluated as to assess the performance in prediction future links between the nodes. The evaluation parameters, also known as metrics, used to measure the performance are accuracy, precision, ROC curve and self-predictability. In this work the proposed method is evaluated with existing ones using accuracy, precision and recall. The Receiver operating characteristics (ROC) curve used to evaluate the prediction methods performance based on entire similarity measure list of the node pair *i.e.* it considers the entire online social network data set [59]. While accuracy and precision emphasize upon the node pairs for which the similarity score is on higher side [60,61]. According to Linyuan Lu, Tau Zhou [62] and Yang et. al.[63] the performance evaluation metrics for link prediction techniques are demonstrated as under.

ROC (Receiver operating characteristics) curve

The ROC curve or area under the curve provides the rank of entire node pairs of the social network between which the links is to be predicted. AUC represents the probability of random selection of link with higher similarity score relative to other non-existing links between the node pairs. Obtaining the sorted list of potential node pairs between which link is supposed to be predicted is quite cumbersome compared to implementation of calculation of the similarity score of unobserved links. Mathematically it is computed as under (Equation 2.18).

$$AUC = \frac{n' + 0.5n''}{n} ---------------------- (2.18)$$

It is to be understood that when obtained value belongs to independent identical distribution then AUC value will be around 0.5, and it shows the performance of link prediction method how better it is predicting the links between then node in a social network.

Precision

Precision referred to the ratio of correctly chosen items to the total number of items selected. Suppose L designated the top predicted links and Lc represents the links that are rightly predicted. The precision would be *Lc/L,* thus accordingly higher the precision value higher would be the accuracy of the predicted links between the node pairs across the social network graph. In terms of classification it is the ratio of truly classified links to the sum of true positive and false positive.

Self-predictability

According to Ciu et. al. described in [59] self-predictability reffered as the measure to evaluate the performance of link prediction techniques. It is measure as per the mathematical expression (Refer equation 2.19) given below and is comparatively less complex than AUC.

$$\delta = \frac{|\ F(G, LB) \cap\ E\ |}{|F(G, LB)|} ---------------------- (2.19)$$

It is to be understood that G represents the social network graph; *E* is the set of links between the nodes in social network graph *G*. Thus *F(G,LB)*is the node pairs set of the networkwhich is predicted using similarity measures larger than the lower bound LB values. δ may be 0 or 1, if $\delta = 1$; shows that node pairs are connected definitely and $\delta = 0$ refers that the social network is unpredictable in nature.

2.8 Approaches, Limitations and Application Areas

The detailed literature survey with features, advantages and limitations of proposed link prediction methods by the various researchers is tabulated below in Table 1. It is a illustration of work performed by researchers from the year 2000. It has been categorized on the basis of prediction parameters, type of network, method used for prediction, application area.

Table 2.1. Network, Parameters, Metrics, prediction techniques along with advantage and disadvantages

Researcher	Link prediction technique	Parameters	Social network dataset	Advantages /Limitations /Novality
Ramesh	Prediction of links	link weights	www, online	No specific advantages,

Sarukkai, 2000 [64]	using markov chains		navigation dataset.	the proposed method is not scalable for large social networking websites.
Huang et. al., 2005	Collaborative filtering based approach for link prediction	Mutual neighbor nodes	Online book store social network	Distance based and mutual neighbor node approaches supersedes the user based and item based techniques
Wang et. al., 2006 [65]	Local probabilistic models and katz link prediction method	Degree of the node	Biochemistry online social networks, DBLP	Classification accuracy is high for dedicated coauthor social networks.
Zhu et. al., 2007[66]	Not specified method for link prediction	Distance between node pair	US patents network	Appropriate only for static social networks(network snapshots)
Yu, K., Chu, W., Yu, S., Tresp, V., & Xu, Z. 2007[31]	Stochastic relational link prediction model and Gaussian process framework	Node attributes	Relational social network	Applied only on relational social network other networks need to be explored
Hasan et. al., 2006	Supervised learning link prediction technique	Proximity distance between nodesand network structural features	Online Social network	No specific advantage/disadvantage

Kashima et. al., 2009[67]	Semi supervised learning link prediction method	Link labels	Multi relational networks	Large number of values stored in main memory for large network thushigh computational complexity.
Murata and Moriyasu, 2007 [32]	Weighted distance proximity similarity measures	Weights of the links between the nodes	Social network	Need to improve the performance of the method in dense online social network.
Song et. al. 2009[37]	Node proximity estimation for social network	No specific parameters found.	Social network	Different types of social network affect the performance of the approach.
Brouard ,Szafranski 2011[33]	Semi supervised penalized output kernel regression	Distance between Node pair	Complex online social network	Use of unlabeled data improves performance slightly.
Lu et. al.,2009 [34]	link prediction based on similarity measure techniques	Local paths between then nodes	Complex networks	Applied only on static social networks.
Narayan et, al., 2011 [48]	De-anonymization technique for predicting future links	Degree of the node s, local attribute	Weighted social network	Performing effectively on weighted social network
Cui et. al., 2016 [59]	Weighted similarity measure based technique to address link prediction problem	Weight of the link between the nodes	Heterogeneous social networks	Method is precisely working on heterogeneous networks. But applied on movie preference social network.

Bliss et al, 2014 [68]	Link evolutionary algorithm for future links	Weight of the link between the nodes	Dynamic online social network, Twitter. Directed social network	Twitter online network data is considered for testing the algorithm.
Kashima, Abe, 2006 [69]	The link prediction problem is addressed using supervised parameterized probabilistic model	Link information as a label is used	Social networks and biological networks specifically directed graph	The proposed method claimed to be effective on directed graphs. Need to explore its effectiveness on other online social networks.
Kunegis and Lommatzch 2009 [49]	learning spectral graph transformations	link weights for similarity computation	Online Social networks	No specific advantage or disadvantage identified.
Fire et. al., 2011 [70]	An ensemble technique is proposed for addressing link prediction problem in social network	Node degree, link subgraph features, path features	No specific social network is used	The proposed method takes more execution time hence time complexity is more.
Lu and Zhou, 2010 [71]	Local similarity indices computed for prediction of missing links	Link weights and local path between the node pairs	Weighted social networks	The algorithm performs poor in case of un-weighted social network.

Lichtenwalter et. al., 2010[39]	Supervised learning used for future link prediction	Distance between nodes	Social network with labels to existing links	The algorithm performance detoriates when labels are unavailable must be explored on networks where labels are not available.
Liu and Li, 2010 [72]	Similarity computing methods	Local random walk	Complex social networks	The prediction accuracy is low.
Bilgic et.al., 2012 [43]	Iterative collective classification	Node labels	Social network	Adv: performed Lim: high link noise reduces the performance.
Chiang et. al., 2011[73	Supervised machine learning based link prediction method	Features derived from longer cycles of the network	Signed social networks	Applied on signed network where relationship is between nodes may be positive or negative. the performance of the method may be explored on other online social networks.
Papadimitriou et. al., 2012 [74]	similarity index and friends recommendation algorithm	Traversing local paths of limited distance	Social network, Hi5, 63 K	The performance of the algorithm decreased when applied on Hi5, 63K dataset.
Liu and Cerpa 2011 [75]	Naïve bayes classifier,Logistic regression, ANN	Link features	Network simulations	computation cost of the algorithm is small. No online social network is

				considered for real time validation
Wang et. al. 2011[76]	Not specified	Human mobility from one place to another	Social network	The accuracy of the proposed method improves mobility information and social ties of humans.
Scellato, Noulas, Mascolo, 2011[77]	Supervised learning framework for link prediction using J48 , naïve byes approach, model trees,linear regression and random forest	Exploits positional or location information	Location based social network	Less space is required improved performance. It performs well on location based social network.
Dong et. al. 2012 [50]	Ranking factor graph model (RFG) and Transfer based RFG)	Distance between potiential node pair across the online social network	Tested and validated on heterogeneous networks	An Improved performance
Feng et.al., 2011 [78]	Clustering perspective model	Local paths between the nodes	Complex social networks	The model performs better in dense social nework. The performance degrades in case of sparse social networks.
Menon and Elkan 2011	Matrix factorization approach for link	Hidden and latent node	Directed Social	Directed social graphs are considered for

[79]	prediction	attributes/features	network	solving link prediction problem. Others networks can be explored
Davis, et. al., 2011[80]	Unsupervised Multi-relational link prediction technique with supervised framework	Weights of the links between the nodes	Heterogeneous information social networks	Supervised frameworks supersedes incase of non trivial. Unsupervised techniques is still domain specific and rigid.
Sarkar et. al., 2012[81]	Nonparametric link prediction	Distance between nodes and temporal information (time stamp)	Dynamic social network	The method includes external features of the graph network.
Barzel and Barabasi 2013[47]	Global silencing of indirect correlations	Path analysis	Biological and bioinformatics social networks	The method translates global co-relation into local information.
Yang et. al., 2014[82]	Community structure based model for link prediction	node degree and distance between nodes for identification of similarity	Complex social network	Method is complex, it may be further studied to reduce the complexity

Jiankun ,Sili 2014[83]	similar information tag and trust algorithm	Path distance and trust of the link	Social network	As the network scales complexity increases.
Guisheng et. al., 2014[84]	Node similarity algorithm based on Link Strength	Similarity measure	Social network	fairly reasonable accuracy with less time complexity. Complexity optimization may be explored
Fu et. al.,2014[85]	Proximity measure approach	Distance between nodes	User item recommender network	Community identification and detection, results vary with nature of the social network.
Aouay et. al., 2014[86]	Technique based on supervised learning (select attribute algorithm)	link labels and degree of the node	Two co-authorship networks	Applied on collaboration social networks. Could be explored on other methods
Ozcan and Gunduz, 2015[87]	Multivariate Time series approach	Link existence temporal information	Evolving dynamic social networks	Not specified
Coskun, Koyuturk, 2015 [88]	Two dimensionality reduction techniques processing sparsity and modularity	Distance between the nodes	Co-authorship networks	Network structure is modular in nature
Shalforoushan &Jalali, 2015 [89]	Bayesian method	Features contributing for	Friendship social networks	Applied on friendship networks only.

		friendship recommendation		
Ahmed and Alkorany, 2015 [90]	Enhanced Friend TNS model	Semantic node(user) attributes, link strengths extracted from the interactions between the nodes	Directed friendship Social network	Applied only on twitter network data set
Malviya , Gupta, 2015 [91]	Evaluation of link prediction metrics	Distance between nodes and node degree	Online social network (Facebook)	Comparative analysis done on Facebook network dataset
Wang et. al., 2016 [45]	Cold-start approach	Latent hidden feature of network structure	Online social network	Not specified
Yang et. al.,2016[92]	Mutual neighbor index based link prediction algorithm	Degree of nodes, information pertaining to community	Social network	Considered local information

De et. al., 2016 [93]	Discriminative method for link prediction	Local features(node attributes)	Community based and Global Signals based networks	Appropriate for moderate density networks.
Zhu et. al., 2016 [94]	Global optimization technique exploiting temporal latent information	Temporal hidden latent space information	Dynamic online Social Networks	Temporal assumption of the information may not possesses relevant information.
Zhang, et. al., 2017 [95]	Link prediction approach exploiting sparse and low rank matrix estimation	Link attributes local link label information	Aligned online social network	Model is effective only for directed social graphs.
Wang et. al., 2017 [64]	ORFP approach without game theory and ORFP-cp approach with game theory	Node pair link weights	Social network	Execution performance affected as number of nodes in a social network increases. net
Wang et. al., 2017 [96]	Algorithm based on node diversity features.	Node Labels and link labels	Social network	Evolution of network based on node diversity.
Zhu et. al. 2017 [97]	Approach based on block coordinate gradient descent	Hidden latent temporal information	Not specified	Optimization of the processing resources particularly time and space.
Wu, J., Zhang, G., & Ren 2017 [98]	Balanced modularity maximization link prediction model (MMLP) a	Degree of nodes in network graph	Online real social network datasets	The approach fairly effective on synthetic network datasets. But it is applicable to only

	community structure based approach.			homogeneous nature of networks.
Moradabadi and Meybodi 2017 [55]	Link prediction approach based on learning Automata	Time sequence of link occurrences between the nodes	Online Social network	The approach is effective over static social network.
Zhang et. al., 2017 [99]	Incremental Dynamic Algorithms	Mutual adjacent nodes(degree)	Social networks	Exploits the relationship information of nearest nodes, in process it takes into the nodes without mutual neighbor nodes leading overhead. Elimination of this overhead may be explored..
Yin et. al., 2017 [100]	Dempster-Shafer theory, entropy of local structure alongwith nodes attribute similarity	Local node attributes along with degree	Social network	Accuracy performance is fairin case of sparse networks only.
Shang et. al., 2016 [101]	Link prediction processing direction information.	Information pertaining to uni-directionand bi-direction	Multi-relation social network	Bidirectional links are informative for link prediction and network structure formation, prediction of links in undirected large network may be explored

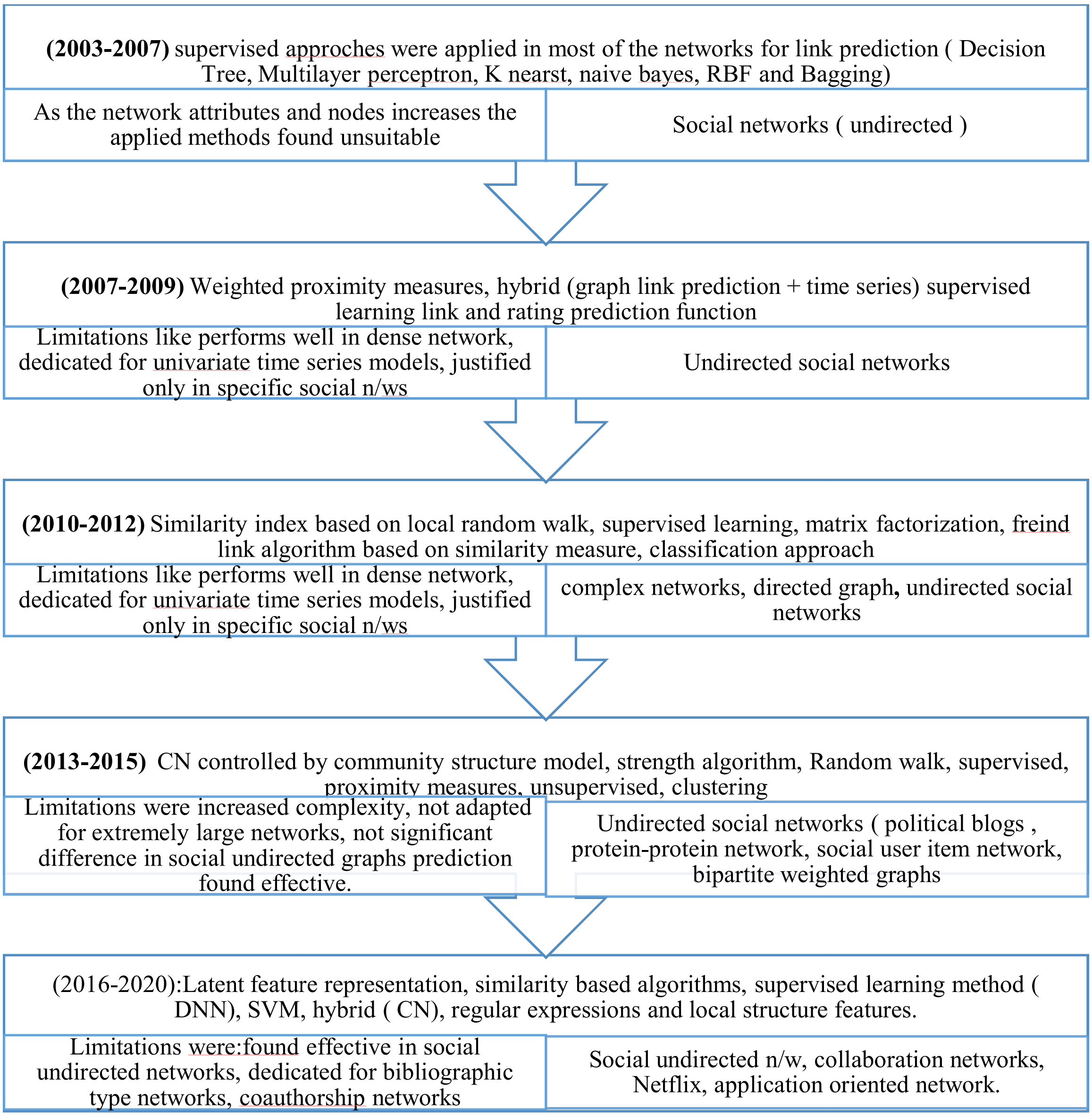
Sharma et.al., 2017 [102]	Link prediction model based on multilevel learning	Consumer preferences	Consumer product websites	Applied on biased and directed network.(recommendation network)
Yao et. al., 2016 [103]	Similarity measure based algorithm	Common neighbors between the node pairs	Online Social network	Enhanced link prediction approach but need to generalized for other online social network.
Srilatha & Manjula, 2016 [104]	No specified Technique	Network structural metrics along with user action.	Online Social network	High computational complexities and overheads.
Ströde et. al., 2016 [105]	Multi-relational link prediction method	Weights of link of node pair	Social network	The disadvantage of the approach is that it bears a high computing cost
Gupta and Sardana, 2016 [106]	Baysian classification prediction model exploiting the local similarity index	Similarity measures	Ego Network	Activity logs other than user profiles may be useful for predictive analysis.
Wang and Bai 2016 [107]	Statistical analysis model and matrix factorization approach	Not specified	Not specified	High variance in performance with differenttypes of network.
Hours et. al., 2016 [108]	Clustering technique for link prediction	tweets andmentions in twitter mention network	Directed social networks(Twitter)	Limited to biased and directed social networks.

Laishram et. al., 2016 [109]	Link prediction approach exploiting link weights	Link weights	Social network	Evaluated on static type of social networks
Amin &Murase , 2016 [110]	node pair link based ranking algorithm	Weights associated to links	Social network	The complexity of the approach is high.
Shu et al., 2017[111]	Deep learning based link prediction approachexploiting similarity indices	Temporal parameters(Time sequence)	Opportunistic sensor network	Applied on opportunistic social networks
Yan and Gregory, 2012 [112]	Link prediction methods based on similarity measures	Community membership information	Not specified	Applied to dense networks only
Li et al., 2018 [113]	Method based on back propagation and Neural network	Meta path parameters/features	Social networks of heterogeneous nature	Applicable to heterogeneous networks
Li et al., 2017 [114]	Iterative updation based link prediction method exploiting link utility	Meta path features	Heterogeneous social networks	deal with heterogeneous social networks only
Shakibian et. al., 2016 [115]	square twin support vector machine link prediction method	Exploiting meta path features as parameter	Heterogeneous social network	Deals with heterogeneous type of links in the network.
	Multilayer complex network	Not specified	Temporal uncertain social network	Remarks: Graph represented using Multilayer complex network

Ahmed et al., 2016 [116]	similarity based on sampling	Time stamps and randomly selected paths	Undirected probabilistic networks	Deals with probabilistic networks and require large computational time and space
Aghabozorgi and Khayyambash , 2018 [117]	Supervised machine learning and mutual node based similarity metric	Social network motifs	Undirected online social network	Simple to implement
Yasami and Sagaei, 2018 [118]	Link prediction approach based on Factorial Hidden Markov model	feature cascade	complex social networks	The proposed method performance may be evaluated on dynamic networks using statistical approach.
Mohan et al., 2017, [119]	Collaborative similarity measure with Adamic/Adar	Local structural attribute	Real world online social networks	Reduced performance due to generating number of messages. The algorithm is implemented onsynchronous parallel programming using Giraph and Graphx(apache).
Dong et al., 2015 [120]	probabilistic link prediction model	Local and global structural features	Social network	Less computational complexity. Combined bi scale information
Bo et al., 2017 [121]	Link prediction model based on trust	Link reputation	Online Social network	The approach impact can be analyzed on the

	traversal and probability	and trust		formulation of links for understanding the performance.
Nguyen-Thi et al., 2015 [122]	Machine learning techniques (Adaboost ,RBF, SVM)	Friends friend and anti-friend whom one dislike	Online Social network	The model needs to be explored on other undirected social networks.
Sherkat et al., 2015,[123]	Ant colony based model optimization technique	Network node features	Social Network	The algorithm supports scalability but tested and validated on a simple network structure.
Shakibian et al., 2017 [124]	Statistical similarity metrics based method.	Path distance(metapath)	Social network of heterogeneous nature	Applied on heterogeneous social networks.
Gao et al., 2017 [125]	Link prediction using projected graph	Node pairs(candidate)	Bipartite social Network	Tested and validated on bipartite network types.
He et al., 2015 [126]	Ensemble hybrid approach using ordered weighted average.	Local network topological information	Social network	Effective predicting links in weighted online social network structures.

After thorough study of existing literature, it has been observed that link prediction problem is relevant presently in prediction of future links in evolving online social networks. Various methods, approaches according to the application areas have been proposed by several researchers. In order to broadly view the approaches along with their limitation and application areas Figure 4 can be referred where we can witness the research done on a particular online social network periodically. Figure 2.9 below can be referred to, in order to generally view the methods along with their weaknesses and implementation areas, where we can regularly witness the research performed on a specific online social network. This diagram demonstrates a broad perspective of link prediction problem undertaken by researcher over the period of time.



(2003-2007) supervised approches were applied in most of the networks for link prediction (Decision Tree, Multilayer perceptron, K nearst, naive bayes, RBF and Bagging)

As the network attributes and nodes increases the applied methods found unsuitable

Social networks (undirected)

(2007-2009) Weighted proximity measures, hybrid (graph link prediction + time series) supervised learning link and rating prediction function

Limitations like performs well in dense network, dedicated for univariate time series models, justified only in specific social n/ws

Undirected social networks

(2010-2012) Similarity index based on local random walk, supervised learning, matrix factorization, freind link algorithm based on similarity measure, classification approach

Limitations like performs well in dense network, dedicated for univariate time series models, justified only in specific social n/ws

complex networks, directed graph, undirected social networks

(2013-2015) CN controlled by community structure model, strength algorithm, Random walk, supervised, proximity measures, unsupervised, clustering

Limitations were increased complexity, not adapted for extremely large networks, not significant difference in social undirected graphs prediction found effective.

Undirected social networks (political blogs , protein-protein network, social user item network, bipartite weighted graphs

(2016-2020):Latent feature representation, similarity based algorithms, supervised learning method (DNN), SVM, hybrid (CN), regular expressions and local structure features.

Limitations were:found effective in social undirected networks, dedicated for bibliographic type networks, coauthorship networks

Social undirected n/w, collaboration networks, Netflix, application oriented network.

Figure 2.9: Illustration of link prediction methods and models proposed by researcher from 2003 onward along with limitations and application areas.

For the prediction of future links in different online social networks, several approaches are suggested based upon features at local and global topological levels. The techniques usually include calculating the similarity measure between the nodes across the entire social network, and it will subsequently used to identify possibility of future relation in between nodes. These similarity measures are of two categories; the first one exploits the characteristics of properties of nodes like the connected adjancent node numbers or mutual common nodes between the node pair [127]; secondly it is approximated by exploiting the networks' structural features like direction, path distance between the node pairs wherein the link is to be predicted [17,128]. The performance of both approaches differs since both characteristics are different from each other, it also depends on the design of the social network, so it is almost impossible to define the characteristics that reliably recognize the similarities between two unconnected nodes and thus predict potential interactions between them. Therefore, the performance of the similarity-based methods can differ in relation to the networks. Different probabilistic models have also been developed to predict future associations or links among nodes by learning from the pro social network to minimize the performance gap [65,81]. The characteristics at local and global level are identified as random variables in these models, and the relationships between them are probably approximated by the Bayesian method through some kind of hidden structure.

Other methods have also been developed in the same way as improving prediction efficiency by recognizing it as a problem of adjacency matrix recovery in which sparse matrix recovery or collaborative filtering has been used to overcome it [79, 129-132]. The statistical models for relation prediction in the social network are beneficial in terms of stability and accuracy [133,134] than the matrix recovery method, according to Zhou et al. in [135], however, on the contrary, the first one is about the matrix recovery method that uses metrics rules as its characteristics and thus emphasizes the topological attributes of the network. Another method is the method used to see the connections that are missing in the direction of rule minimization,[79,136] and it occurs that the actual network structure may not act appropriately and is therefore directly represented in the representations of the matrix.

As per Oczan and Oguducu, [137], The traditional node similarity techniques can fairly used for solving link prediction problem, in their research they suggested a multivariate NARX model that predicts links along with the recursively occurring of existing links in heterogeneous social networks [87]. To fix the issue of relation prediction, they have used neural network strategies. In addition, the work on networks with several types of nodes with links that shift over time has been expanded. The suggested model is based on VAR models that are essentially introduced for social networks with multivariate time series information. Li et. al described in [57] that a supervised link prediction model was also proposed supervised link prediction model for prediction of possible links in complex heterogeneous networks. They exploited the network for extracting meta path features to build a neural network model to predict a link between the node pairs. The authors have endorsed that the method supersedes in comparison with the classical baseline methods specifically in heterogeneous networks.

In [138], Mallek et. al has introduced a noval link prediction model that considers uncertainty of the link existence between the node pair. They have used belief function theory as to quantiyfy the existence of the links. Now in order to extract the evidential information about the link existence they used neighbor node features. As per Aghabozorgi and Khayyambasi [117], the structural entities of social network graph along with their attributes may be used to compute the similarity measure for recognition of node pairs between which link existence probability estimated. They evolved a similarity metric exploiting the network motifs and validated the method using supervised machine learning framework.

2.9 Social network data formats and analysis

Most of the studies carried out so far indicate that analysis of the network is either direct or indirect method of graph analysis, or it may be assumed that these networks are described as graphs and are extensively studied from a topological point of view. The graph properties studied include size of graph, its density, degree distribution of nodes across the graph, path distance between graph nodes, coefficient of clustering, common neighbor node, network communities, potential nodes and links between them, etc. As described by Faloutsos et. al. in [139] distribution of degree on the internet usually in line with the power low which is also having its evidences in graph theory demonstrated for www [140,141]. Similarly the friendship networks and email networks been analyzed to understand the user interaction or node interaction across the entire online social networks [142]. It has been observed while study that the analysis done so far has been carried out on static network graphs on the contrary present online social networks are evolving exponential due to availability of internet. It may be referred in the latest reviews carried out by researchers in [143-145].Snapshots of such networks are taken and analyzed at periodic times in order to examine the real world data of such social networks such as Facebook, Twitter. Fetterly et. al. [146] and Chao et al. [147] has illustrated the topological properties of various social networks snapshots.

2.9.1 Types online social network datasets for analysis

Around 30 to 40 years ago, massive data had been stored from several large online social networks; at that time sufficient software and hardware were unavailable as to crunch the network data appropriately and accordingly. Technology has progressed and evolved to a degree that the systems are now capable to manage the whole lot of large network datasets. Python with rich API pool is capable of simulating, creating and leveraging such networks in compliance with society's requirement for further decision-making and prediction. Thus, in the suggestion of potential nodes and link between the nodes, the exploitation of structural and local features extracted from the online network portion can be of benefit.

Many software tools are available to process the large and complex datasets of the social networks, like Python, R, Gephi, etc. The online social network datasets are also present in various formats according to the software available that can be selected according to the tool to be used for processing. Following are the common formats of online social networks. These online social network dataset formats can be downloaded from https://snap.stanford.edu/data/.

1. .CSV
2. .GML
3. .Pajek Net
4. .GraphML
5. .GEXF

2.9.1.1 Comma Separated Values (.CSV)

CSV format contain social network structural information about the nodes and links. The file typically contains columns and rows, containing node source and destination node identification numbers, along with the weight of the links between nodes. The dataset can be accessed in the .txt or .csv extension formats as an edge list of adjacency lists comprising node to node associations.

2.9.1.2 Graph Modeling Language (GML)

In this format the node and link information of the network is available in text mode. Using python with network could be appropriate for any GML record. The instructions for accessing and crunching the GML social networks are as follows:

Pseudocode:

```
g1 = nx.path_graph(10)
nx1.write_gml(g, 'test.gml')
r1 = nx.read_gml('test.gml')
```

2.9.1.3 Pajek Net Format

Social network datasets available in pajek format need .NET augmentation. For implementing a chart using pajeck the dataset needed to be provided in the form of string storing diagram G. the output will be a multi graph of multidigraph. The instructions for implementing the functions using python's networkx library is as follows.

```
g1 = nx.path_graph(10)
nx.write_pajek(g, "test.net")
g1 = nx.read_pajek("test.net")
```

In case if a multigraph , follow.

```
G1= nx.Graph(G)
```

2.9.1.4 GraphML Format

This format is simple and comprehensive file format containing the information of graph. It used to store the structural and topological attributes of graph and further extendable when more information like application particular data is supposed to be stored. It supports all types of graphs such as undirected, directed, hybrid, hierarchical, references to external data and application specific data. The format is based upon XML and does not supposed to use customized syntax. Therefore it is appropriate for all types of processing. To read graphML format it requires a path string is passed as a parameter which returns a digraph or a graph.

2.9.1.5 GEXF: Graph Exchange XML Format

Graph Exchange XML Format illustrates the structure of large complex networks or complex graphs by storing pertaining data and dynamics. It was developed with Gephi project which has been extensively working on issues related to graph exchange and development processes. Since then its specifications are now fairly seasoned, extensible and appropriate for real time applications like online social network analysis. Mathematical illustration of graph structures is quite simple and possible using this. A string of the path will be passed as a parameter to the function returning graph, digraph, multigraphs or multidigraph.

```
g=nx.path_graph(10)
nx.write_gexf(g, "test.gexf")
```

2.9.1.6 Dataset Format Summary

Many graph formats and module are available for simulating and experimental study of social network analysis. These tools need certain functions that are used for processing the graph datasets. The programming technologies like python have a rich library of packages that supports external files and other programming utilities. Table 2.2 illustrates the reading and writin procedure to exploit the various common network graph dataset formats [148,149].

Table 2.2 Corresponding procedures and functions for crunching online social network datasets.

Dataset File Format	Short name	Reading Procedure	Writing Procedure
Adjacency list	Lgl	Graph.Read_Lgl()	Graph.write_lgl()
Adjacency matrix	Adjacency	Graph.Read_Adjacency()	Graph.write_adjacency()
DIMACS	Dimacs	Graph.Read_DIMACS()	Graph.write_dimacs()
Edge list	edgelist, edges, edge	Graph.Read_Edgelist()	Graph.write_edgelist()

GraphViz	graphviz, dot	not supported yet	Graph.write_dot()
GML	Gml	Graph.Read_GML()	Graph.write_gml()
GraphML	Graphml	Graph.Read_GraphML()	Graph.write_graphml()
GzippedGraphML	Graphmlz	Graph.Read_GraphMLz()	Graph.write_graphmlz()
Labeled edgelist	Ncol	Graph.Read_Ncol()	Graph.write_ncol()
Pajek format	pajek, net	Graph.Read_Pajek()	Graph.write_pajek()
Pickled graph	Pickle	Graph.Read_Pickle()	Graph.write_pickle()

2.10 Summary

In this chapter we have thoroughly discussed the literature available on link prediction problem in social networks. The chapter also describes the background of online social network, how it has been evolved, how it can be represented in the form of sociographs. Furthermore the chapter also explains the various link prediction metrics available for measuring similarity between the nodes in social network along with various other methods and models proposed by researchers.

Chapter 3

Methods and Material

3.1 Overview

Previous chapter has illustrated the literature study of social network analysis and link prediction in online social networks. The representation of social network as graphs, components of graphs and its structural entities were also described in detail. Based on the critical analysis of the link prediction problem and the works proposed by the earlier researchers there are certain research issues which have been drawn, discussed in chapter 1. In this chapter 3, the methodology adopted for addressing the link prediction problem along with the description of algorithms used for extracting the features from the online social network is discussed. The feature extracted will be used in further classification and determination of potential node pair between which the link is supposed to be predicted. The method described is bifurcated in two modules: the first module is to compute the measure on the basis of which similarity between the node pairs identified. Secondly the network formed with the features will subsequently process for future link classification. Two approaches have been described here; a collaborative clustering-based approach or model is proposed in which topological attributes are exploited which will be used in further link prediction. In this model clustering-based supervised machine learning approach is employed that converges the network such that the final subsequent network obtained has the node pairs between which the possibility of link prediction will be strong. The method is based on shift k-means approach. In addition to this the chapter also describes proposed probabilistic link prediction techniques which compute the measure that contributes to the probability of link existence between the node pairs across the online social network. It also presents the overview of the machine learning techniques used for the analysis of the social network datasets with extracted measures. The initial section of the chapter gives an overview of the field of social network analysis and statistics. The chapter also includes the introduction about the software's tools used to store and analyze network datasets.

3.2 Online network data format

The structure of online social network is represented through graphs and also referred as sociographs. The users of online social networks (Facebook, LinkedIn, etc.) are represented through nodes and the relationship between these nodes is represented through links. The information of graphs is stored in the form of edge list or adjacency list. In our research the network dataset used is in the form of edge list. An example of online social network data set is shown below; the sample is taken from Facebook, Github, Twitch and Deezer network dataset edge list, in this the first node maps to the second node (Refer Table 3.1):

Table 3.1: Sample social network data set formats

Deezer		**Facebook**		**GitHub**		**Twitch**	
84	152	1535	0	7740	20363	3243	1916
1469	4802	1536	0	34840	20363	1006	1860
4802	1469	1537	0	27367	20363	1451	1860
19701	1469	1538	0	6073	0	3080	1860
14092	1469	1539	0	11635	896	2794	1860
13484	1469	1540	0	896	11635	1175	1860
9736	1469	1541	0	32126	11635	1916	1800
24954	6268	1542	0	4382	11635	1244	1860

The data set consist of two columns separated by space having number of rows. The first column contains name of the node in the form of ids that has a connection with the corresponding node in the second column. The connection represents a relationship or interaction between both the nodes. The datasets have been used in [150-152].

3.3. Social Network feature extractions

The dataset described in above section is the base which is further considered by researchers to perform experimental calculations for extracting the features. The features extracted are typically saved in csv a format which is a comma separated social network data with attributes or features. An entire row in the excel file contains the node pairs and their related attribute values which in our case is similarity measure calculated on the basis of node neighborhood property. In some researches it is also known as metric. In the terms of machine learning respective rows and columns of the network data referred as instances and attributes [153]. There are tools and technologies available to test and analyze the social network datasets to address the link prediction problem. In Our research python has been used for computation of similarity measure between node pairs and Weka for applying machine learning techniques for further classification analysis.

3.4. Performance evaluation

The evaluation of link prediction model is performed on the basis of how accurately it is predicting the links between the node pairs. While evaluating machine learning methods the large network data set is divided in to training, testing and validating dataset [153]. The performance of the approach calculated how closely it is classifying the legitimate links close to 100%.

Accuracy: It is actually the measurement of closeness to the legitimate existence of the link predicted between the node pair. It can be defines as the description of systematic errors, a statistical bias measurement; high accuracy means the difference between the true approximation and the result is very less and low accuracy refers the difference is comparatively high. This means there is no correlation with the class types.

Precision: It is illustration of random errors, a statistical variability measurement. It evaluates the fraction of correct classification of instances amongst positive classified values. Thus according to the formula it is: TP/(TP+FP)

The calculation of performance parameters depends upon following.

True positive(TP): True positive classification refers to the case in which the model predicts the future links correctly *i.e.* 'yes' that means there is actually a future link between the nodes and the modes has predicted it correctly.

True negative (TN): True negative classification refers to the case where the model has predicted that there is no future link between the node pairs and prediction is 'No'.

False positive (FP): It refers to the mis-prediction of the future link, our model has predicted 'yes' there exists a future link, but in fact there is no link exists between the nodes. Sometime it is known as type I error.

False negative (FN): False negative is just reverse of true negative. While prediction of future links the model has predicted 'No' there will be no link exists between the node pairs, but actually there exists a link. Sometimes referred as type II error.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} - - - - - - - - - - - - - - - - -(3.1)$$

$$Precision = \frac{(TP)}{(TP + FP)} -(3.2)$$

$$Recall = \frac{(TP)}{(TP + FN)} - (3.3)$$

Alternatively accuracy can also be evaluated using the kappa statistics defines as PFA-PE/1-PE, where P(A) represents the percentage accuracy of links predicted by the learning model whereas P(E) is the percentage of accurate predictions made through random guess [154]. The range of kappa is from -100% to 100% and usually denotes how useful the model will be as compared to the random guess [155]. Nevertheless it will not consider the true positive increased values, usually the values more than 40% shows that there will be reasonable chances to have a link between the node pairs in online social network. In addressing link prediction problem this statistics is not going to be used significantly for evaluation of the link prediction techniques and methods.

3.5 Research Methodology adapted

In our research an algorithmic research methodology to address the link prediction problem in online social networks have been performed. The steps which have been followed for doing the research work are as under.

1) The evolution of links prediction particularly along with an impact of machine learning techniques would be studied.

2) Thoroughly go through the distinguished basics of evolution of social networks in order to classify them in directed and undirected graphs.
3) Different link prediction schemes would be analyzed and applied over the available dataset and studied from the point of performance, it will help us in identification & devising a method to predict links between the nodes by considering the node attributes.
4) A study of exactly the available prediction techniques applied for link prediction for identifying the missing links amongst the nodes would be done and subsequently a model would be proposed.
5) Implementation and simulation of the method using open source.
6) Developed model would be tested to validate and verify the performance using area under the ROC (Receiver operating characteristics) curve (AUC).
7) The results of proposed link prediction approach for social networks would be analyzed and evaluated.

3.6 Proposed probabilistic link prediction method

Thorough study of online social network evolution has inferred that the growth of the social network often displays certain patterns to understand. The attributes of nodes and links contribute significantly in forecasting the estimation of new links between the nodes across the network. The proposed model of link prediction is demonstrated in Figure 3.1. In our method the node neighborhood properties have been exploited to compute the probability contribution in prediction of link between the nodes.

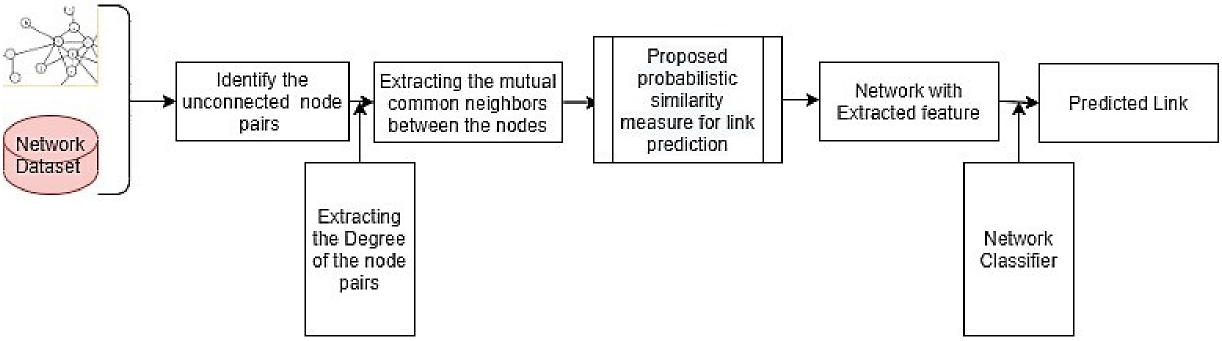

Network Dataset
Identify the unconnected node pairs
Extracting the mutual common neighbors between the nodes
Extracting the Degree of the node pairs
Proposed probabilistic similarity measure for link prediction
Network with Extracted feature
Network Classifier
Predicted Link

Figure.3.1. Process to predict future links between the node pairs in a large and complex online social network

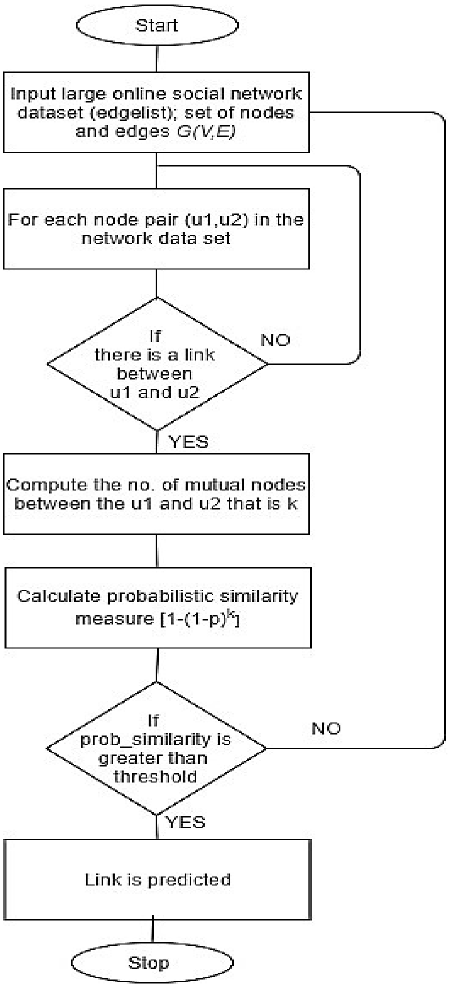

Start
Input large online social network dataset (edgelist); set of nodes and edges G(V,E)
For each node pair (u1,u2) in the network data set
If there is a link between u1 and u2
NO
YES
Compute the no. of mutual nodes between the u1 and u2 that is k
Calculate probabilistic similarity measure [1-(1-p)^k]
If prob_similarity is greater than threshold
NO
YES
Link is predicted
Stop

Figure3.2. The flow diagram of predicting a future link between node pair in large online social networks.

3.6.1 Probabilistic Similarity Measure Algorithm

According to the method assume that the probability contribution of having a link between node pair is computed as.

Let '*p*' is the probability of two nodes to get connected when having 1 common neighbor node. Then the probability of not getting connected in future within the network mathematically will be computed as:

$$1-p - - - - - - - - - - (3.4)$$

Approximating the probability contribution for having a link between node pair where there are n number of common nodes, therefore now for 'n' number of common nodes the above mathematical equation may be written as:

$$(1-p)^{n} - (3.5)$$

Thus, the probability of having a link between the nodes pair having n number of neighbor nodes is determined by subtracting the equation no (10) from 1, therefore the equation of probability similarity measure for prediction of link between the node pairs for n number of neighbor nodes across the network can be written as.

$$Prob_{sim}.(u,v) = 1-(1-p)^{n} - - - - - - - - - - (3.6)$$

Where 'n' is number of common neighbor nodes between the node pair u and v in an online social network

The algorithm.

```
1. Input node u and v between which link is to be predicted, such that there is no link in
   between.
2. For each node pair
      I.   Compute the degree of node x and node y
3. For each node pair where node u! =v
      I.   Compute 'n', the common neighbor nodes between nodes x-y in the network.
      II.  Compute:      $prob\_sim=1-(1-p)^{n}$
4. Initialize the value for p such that it should lie between the range 1 to 0.
5. If prob_sim > threshold value

      I.   Predict link between x and y
```

3.7 Proposed collaborative clustering approach

Link prediction in social network is a problem of social network analysis, presently the online social network applications has been generating data exponentially in volumes. Such large volume social network data is quite complex to analyze, machine learning approaches has significantly and effectively helps in processing such large and complex data. A machine learning based clustering approach has been proposed to process the social network data that used to classify into groups or subgroups termed as clusters. As discussed clustering extracts relevant information out of the dataset that helps in grouping and it is effectively and efficiently applied to process voluminous data of different nature including social networks and business research. Clustering techniques conceptually computes distance in between the objects also referred as Euclidian distance. The difference of distance later used majorly for creating clusters. In our collaborative clustering approach, it processes the online social networks in two folds. The first part captures the social networks data and extracts the topological features exploiting the mutual node neighborhood property of the network, thus extracted features will be used for computing the similarity measures in order to judge the proximity between the nodes for predicting future links. Social networks are usually represented through graphs containing nodes and links between nodes. As far as network features are considered, classified as local features (node attributes) and global features (distance between the not connected nodes) and our research we considered local features. With new information the ordered node pairs of social network between which potential links is to be predicted is further analyzed using shift k-means clustering technique. That way the process will reduce the size of the large and complex online social network to relatively small social network graph. The features extracted using proposed probabilistic link prediction algorithm. Existing topological link prediction techniques like (JC, AA, CN and PA) can also be used to compute the topological local features of the network.

3.7.1Clustering Algorithm

The proposed link prediction algorithm is based on unsupervised learning, which is used to converge the online social network data thereby enhancing possibility and probability of future links between potential node pairs. Consider a social network graph data comprising set of nodes V such that $V = (v_1,v_2,v_3,...v_n)$.Assuming the another set X such that $X = (x_1, x_2, x_3 .... x_n)$, X is the set of dimensional vector values obtained for the node pairs between which links is to be predicted. The purpose of the algorithm is to divide the obtained values in a k clusters such that $k <= n$ like set $S = \{s_1, s_2, s_3, ..., s_k.\}$[156,157]. In other words, to summarize the process the algorithm starts with selection of the number of required subgroups or clusters and initialize the centroids (vectors calculated of potential node pair) of the groups, which is actually an approximated value out of calculated similarity measures for the node pairs between which link do not exists.The closeness distance of each corresponding node pair is identified from the centroid and assigning it to the group having nearest centroid value. Accordingly, the node pairs will be assigned to the cluster. In the same fashion all the data values of the node pairs will be processed and accordingly clusters will be identified. After first iteration, same process will be followed in the subsequent iterations and centroids for the clusters are recalculated. The entire repetitive process continuous until the same centroid values obtained along with the same set of values . Figure 3.1 demonstrates the flow diagram of the algorithmic process in detail.

The recursive computation of closeness distance of specific node pair similarity feature value from the centroids is done using a mathematical equation represented in (3.7).

$$\arg\min(S) = \sum_{i=1}^{k} \sum_{x \in s_i} |(x_i - c_i)|^2 \tag{3.7}$$

It is to be understood that , c_i is the centroid value of a cluster set (S_{i})

The Algorithm

Let x is the set of similarity values determined by exploiting the node neighborhood properties of the node pair for which the link is to be predicted; such that $x = \{x_1, x_2, ...x_n\}$. also assume a set V containing mean centroid values such that $V=\{v_1, v_2, ...v_c\}$

Step 1. Select the centroid 'c' of cluster group randomly

Step 2. Calculating closeness of the cluster data elements from the centroid by obtaining the distance'd'.

Step 3. If the observed difference is relatively less than the other data element closest to the cetroid. Assigning the node pair to the cluster where the distance from the centroid is minimum.

Step 4. Iteratively repeat the same method for all cluster centroids.

Step 5. Recalculating the centroid of the clusters; such that c_i is the no. of data elements in the i_{th} cluster.

Step 6. Recalculating the closeness of data elements from newly obtained centroid by determining the distance from the centroid and reassigning the node pairs accordingly as to get new cluster.

Step 7. If new data elements or node pairs are not being identified for assigning

Step 8. stop

Step 9. Else

Step 10. Repeat the process from step no. 3.

Step 11. End

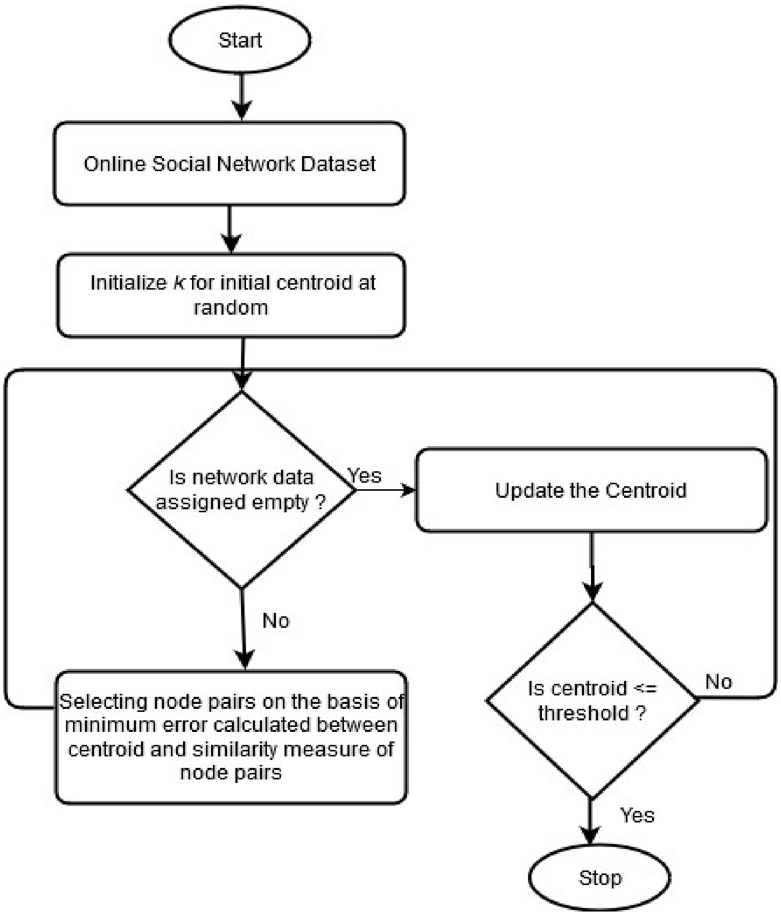

Start
Online Social Network Dataset
Initialize *k* for initial centroid at random
Is network data assigned empty ?
Yes
Update the Centroid
No
Selecting node pairs on the basis of minimum error calculated between centroid and similarity measure of node pairs
Is centroid <= threshold ?
No
Yes
Stop

Figure 3.3. Clustering process for link prediction

3.7.2 Collaborative link prediction model

The proposed collaborative links prediction model is developed to exploit the benefits of two distinguished link prediction techniques in online social networks. The model works in two modules namely similarity feature extraction module and link classification module. In similarity feature extraction module topological mutual node neighborhood properties exploited for identification of similarity on the basis of which proximity between the nodes identified. The similarity techniques used for computing the similarity measures are proposed probabilities similarity measure along with JC, AA and PA. The process is applied on the node pairs between which there is no link available in current state of the social network snapshot. On the basis of obtained similarity measure all the potential node pairs are sorted in ascending order. The higher the score of the node pair, it is very likely to have a link between such node pairs. These obtained similarity values helps in further classification of legitimate link between the nodes using machine learning approach. The classification model uses *k*-means clustering technique for converging the network into appropriate clusters of potential node pairs between which link is to be predicted. The benefit of *k*-Means clustering technique is that it is simple, fast and will not consume much time for processing the entire social network. Figure 3.2 demonstrates the methodology in detail.

As known the online social network data set is voluminous and complex carrying different feature, for building a link prediction model therefore the network dataset for further processing is needed to be prepared. As discussed for creation of clusters similarity measure of the node pairs are used as a feature, total number of clusters (k=2) has to be provided to the algorithm for further crunching the network. The recursive iterations of measuring the closeness distances repetitively is in this case uses similarity measure instead of Euclidian distance. The accuracy of cluster classification subsequently improves with every iteration. The value of *k* here in this algorithm is quite significant and should be selected wisely, such that the clusters should be dense and prediction of links between the node pairs will be accurate and precise.

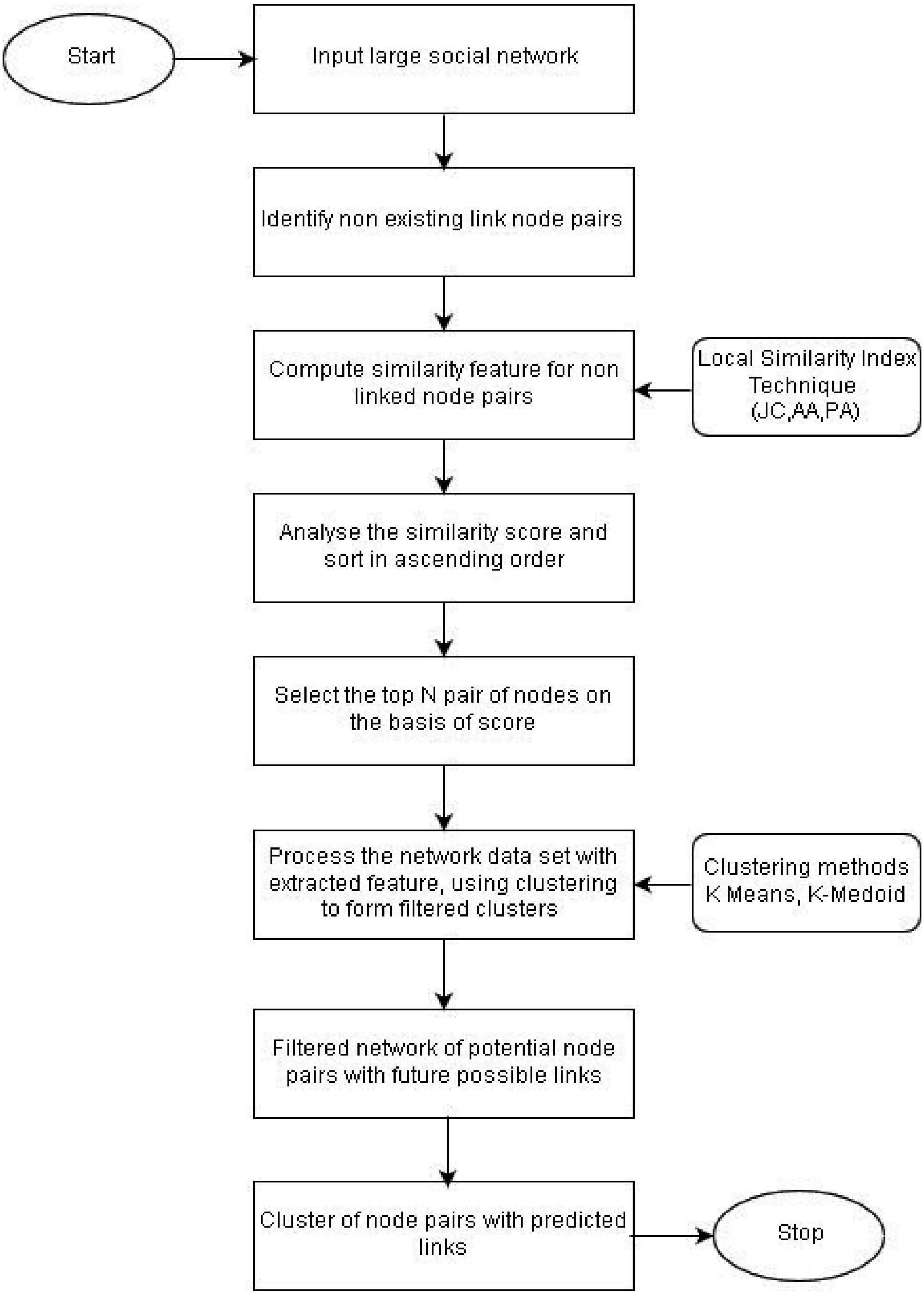

Start
Input large social network
Identify non existing link node pairs
Compute similarity feature for non linked node pairs
Local Similarity Index Technique (JC,AA,PA)
Analyse the similarity score and sort in ascending order
Select the top N pair of nodes on the basis of score
Process the network data set with extracted feature, using clustering to form filtered clusters
Clustering methods K Means, K-Medoid
Filtered network of potential node pairs with future possible links
Cluster of node pairs with predicted links
Stop

Figure 3.4. The ensemble collaborative model for link prediction

Algorithm

Step 1. Input the network data edge list, *G* (*u,v*).

Step 2. Identifying node pairs (u_1, u_2, u_{n-1}) between which there is no link is available.

Step 3. For each node pair in graph *G,* calculate the number of mutual neighborhood nodes between node *u* and node *v* by computing.

$$\Gamma(x) \cap \Gamma(y)$$

Step 4. For each node pair in graph *G* calculate the similarity value by computing.

$$\text{Similarity value} = p-(1-p)^n$$

Step 5. Arrange all the potential node pairs in descending order from *1* to *n-1*, and select the top N number of node pairs.

Step 6. Assign the value of k to define the number of clusters and select randomly an arbitrary centroid '*c*' of clusters.

Step 7. Calculate the closeness between the centroid and the similarity values by computing the difference between them.

Step 8. Comparing the difference between the cluster centroids and the similarity value and assign the node pair to the cluster where the difference with the cluster centroid is relatively less then the other.

Step 8. Apply the same process until the last node pair is processed and assigned to appropriate cluster.

Step 9. Repeat the process by recalculating the cluster centroids again for all the identified cluster as to optimize them.

Step 10. If the calculated obtained cluster centroid is new than

Step 11. For each node pair again determine the closeness distance in between the similarity measure and new centroid, reshuffle the cluster by assigning the node pairs accordingly to the cluster such that the obtained closeness distance is minimum.

else

Step 12. Stop

Step 13. Repeat the algorithm until no new centroid and node pair is identified for further allotment.

In this process it is to be understood that the repetition of the entire process of converging the social network dataset to relatively small dense sub graph will be terminated when local minimum similarity value is achieved. It will be terminated as soon as the error obtained between selected centroid and the similarity measure of the node pair comes below the defined threshold value. The error is calculated using equation mentioned below:

$$Error\ (correction) = \sum_{i=1}^{k} \min_{i=1} \left|\left|x_i - c_i\right|\right|^2 \tag{3.8}$$

Where x_i is the similarity value between the node pairs; c_i is the selected centroid.

3.8 Summary

In this chapter the methodology used in addressing the link prediction problem in online social networks is discussed. Further it has illustrated the two methods for link prediction; an algorithm that computes the similarity measure which contributes in the probability of predicting links between node pairs across the entire social network, this method exploits topological local attributes of the network structure for obtaining the similarity measure. The obtained similarity measure helps in approximation of future links between the nodes. Introduced a supervised shift *k*-means machine learning based model that converges the large online social networks in the cluster of nodes between which the probability of link existence is high.

Chapter 4

Experimental Analysis

4.1 Overview

In the previous chapter the methodology used in addressing the link prediction problem in online social networks is illustrated. This chapter describes the experimental environment required for implementing the methods proposed in previous chapter. In this chapter the data sources along with the description of datasets used in the research work is described. Moreover the programming technology and tools used for online social network data analysis is also being discussed. Python has been used for implementing the similarity algorithm and Weka is used for applying the machine learning method. Moreover the results obtained after processing the real time datasets has been illustrated and evaluated using state of the art performance parameters.

4.2 Experimental environment

The experiments for the justification of link prediction and its techniques in online social networks have been carried out using programming and data analysis tools. Simulations of networks are being carried out using python 3.7. Networkx package is extensively used for generation of networks, nodes and edges across the network. For further social network data analysis Waikato environment for knowledge analysis (WEKA 3.9.9) has been used. The tool is an analysis tool developed by university of Waikato. It is an efficient tool to apply data mining functions such as data preprocessing, classification, clustering, regression, feature selection and visualization. It has been developed in java language.

The datasets used for training, testing and verification of the methodology are Facebook, Wikipedia, Deezer, GitHub and Twitch online social networks.

The Facebook dataset consists of edge list between (friends list) extracted from Facebook. The data was gathered from the survey undertaken on Facebook app. It consists of edge list between users. The data captured has user ids replaced with anonymous one. For example, the original dataset may carry a resemblance "political=Democratic Party" the new name for this would be "political=changed feature 1". It will hide the actual affiliation of the user;however, the implementation of link prediction problem is not going to be hampered with this change.

4.3 Dataset information

In this research work for experimental analysis five online social network datasets are considered for applying the link prediction methods namely; Wikipedia, Deezer, Facebook, Twitch, and GitHub. All these online social network datasets are scrapped using social network data set website: https://snap.stanford.edu/data/ and described as under.

Wikipedia

To understand the performance of the approaches designed for approximation of future links between the node pairs which are not connected with each other in an online social network, Wikipedia dataset has been used. Wikipedia is an active community of users who writes articles of interest on the Wikipedia wall. The network is being studied to analyze the votes given by users to choose the admin. The network is a kind of signed network where the users voted on the promotion of another. A vote in support represented by positive vote and one in oppose would be represented by a negative vote.

Dataset crawled contains vote history, containing around 2,794 elections referred ad nodes along with 103,747 votes cast amongst 7,118 users, represented as links between the nodes, which includes existing admins. As mentioned this too downloaded from https://snap.stanford.edu/data/wiki-Vote.html. The nodes of the wiki network are the users and directed links are present from node '*i*' to '*j*', it comprehends that the designated user '*i*' has voted user '*j*'.

Deezer Social Network

A Deezer social network that was collected in March 2020 from the public API. Nodes are users of deezers from countries in Europe and their connections are common followers. The features of the vertex are derived from the artists the users prefer. The role of the graph is to predict the gender of the consumer in binary node classification. The target feature each user has been extracted from the name field [150].

Facebook Large Page-Page Network

This is a page graph of checked Facebook websites. Official Facebook pages are nodes, while ties or links in between sites are mutual. The node features are taken from the site descriptions created for the function of the site by the page owners. This dataset was collected in November 2017 via the Facebook Graph API and restricted to Facebook-defined pages from four categories. Which are political, governmental, television and business types. .The task for this dataset is to identify multi class nodes for the 4 categories of the site [151].

GitHub Social Network

The downloaded GitHub dataset is a large social network community of GitHub developers collected using the public API in June 2019. The developers are represented by nodes, who shared around 10 repositories and links represents the mutual followers interaction or association amongst them. The node features are fetched on the basis of thier localtion, the repositories shared, the employer and the email address. The task that can be related to this graph data set could be binary classification where one can predict a GitHub user is a machine learning developer or a web developer. The target feature was extracted from job title of the users [151].

Twitch Social Networks

The scrapped Twitch social networks dataset is usually used for transfer learning and classification of nodes. The twitch user networks are basically gamers who used to stream in specific languages. It is obvious that the users are the nodes and the links are the mutual friendship amongst them. The node features are usually fetched on the basis of the game played and liked, the location from where they used to play and the streaming habits of the user.The dataset actually shares the same set of node feature which makes transfer learning across the online social network possible. The social network dataset is collected in May 2018. The task which can be related to this social network data could be binary classification of node wherein it can be predicted that whether a streamer or gamer uses a explicit language or not [151]. The statistics of all the online social network dataset used in the study are tabulated in Table 4.1.

Table 4.1. Online social network dataset description and statistics

Dataset Properties	Twitch Social Network	Facebook social network	Github Social Network	Deezer Social Network
Nodes	4385	22470	37700	28281
Edges	37304	171002	289003	92752
Density	.004	0.001	0.001	0.0001

The selected online social networks are equally dense except twitch which is slightly denser as compared to others, which means it has less number of nodes and in between connection than the former social networks.

4.4. Result and discussion

The 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 results obtained using the algorithm are shown in following sections.

4.4.1 Probabilistic Similarity Measure

The proposed technique is implemented using python. Initially A hundred social network has been created to with random edges between the nodes just to understand the evolution of the network with respect to time. The local properties of the network are processed repetitively leading calculation of similarity measure between the node pairs. The probabilistic link prediction technique is applied over four different online social networks for further evaluation on the scale of accuracy, precision and recall. The obtained network dataset is then applied with machine learning classification techniques.Further classification of legitimate links predicted between the node pairs the obtained network data with features is processed on weka.

The observed experimental developments in the network have been demonstrated in Figure 4.1(a) and (b). Figure 4.1(a) illustrates social network obtained at time ‘t’ and Figure 4.1(b) is the illustration of subsequent network which is dynamically generated by calculating the probability similarity measure between unconnected node pairs u and v. It is visible from the relative comparison that new links has been formed between nodes across the entire network.

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 4.2.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 4.2. 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

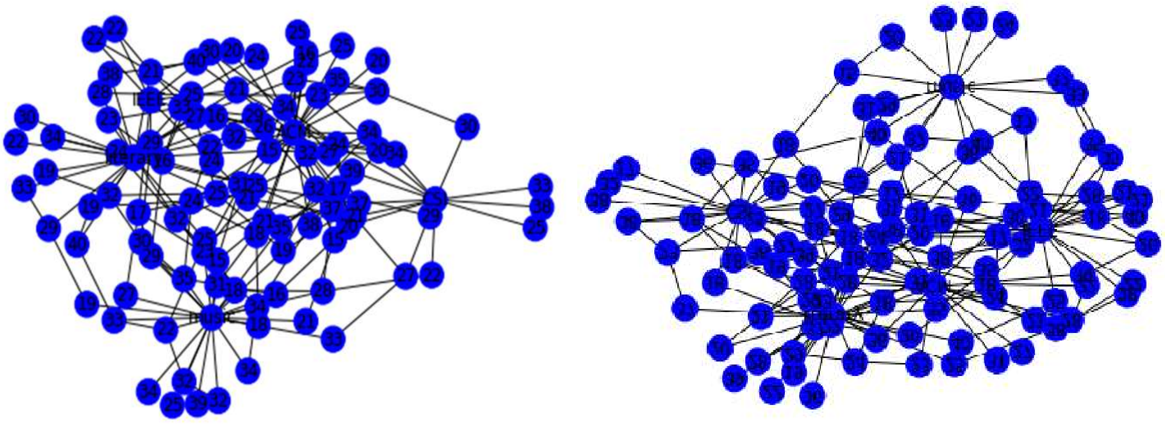


Figure 4.1. (a) Evolution of social network at time‘t’; (b) The graph obtained at time‘t+1’

4.4.2 Comparative analysis

Experimental analysis carried for evaluation of proposed technique with state-of-the-art link prediction techniques. Classifier built using machine learning methods trained with the social networking data set along with extracted features from node properties. The link prediction observations obtained by applying the algorithm on Twitch, Facebook and Deezer online social networks are tabulated below. For better understanding about the performance of the proposed algorithm it has been applied over all the online social network datasets. Table 4.3 shows the comparative analysis of proposed similarity algorithm on Twitch online social network. Similarly, Table 4.4, Table 4.5 and Table 4.6 are demonstration of Facebook, Deezer and GitHub online social networks respectively.

Table 4.3. 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 4.4. 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 4.5. 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

According to the results obtained it is clearly visible that the probabilistic similarity technique is contributing fairly good in comparison with the other state of the art link prediction techniques. The results obtained for Jaccard coefficient it is quite unrealistic to obtain nearly 100% result, this may be due to the size of the dataset and the features extracted. The most of the similarity feature obtained for JC are similar for most of the node pairs. Therefore, in this respect we can ignore it.

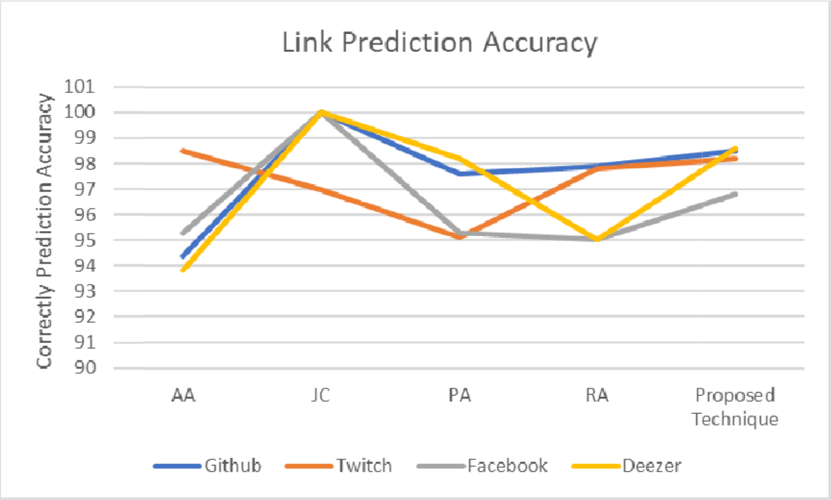


Figure 4.2. Relative comparison of correctly classified link prediction accuracy across Facebook, Twitch, Deezer and GitHub networks.

The relative comparison of correctly predicted the future links between the node pairs in the social network has shown that the proposed probabilistic method is significantly contributing in addressing the link prediction problem. Figure 4.2 demonstrates the accuracy analysis of the proposed method with the state-of-the-art link prediction techniques across the selected online social networks: Facebook, Twitch, Deezer and GitHub.

4.4.3 Collaborative clustering approach

This method is based on *k*-means clustering approach, in our version of algorithm, classical similarity measures have been used instead of Euclidian distance which is obtained using method JC, AA, and PA. The results obtained after applying these techniques are tabulated in

Table 4.7. The experiments carried out on real social network datasets (Wikipedia, dataset described above) it is also downloaded from https://snap.stanford.edu/data [158].

The dataset used for analysis does not contain time stamping of link formation or between the node pairs across the network at time t. Now It is difficult for us to segregate the network in two parts, one is at time *'t'* and second is at time *'t+1'*. In order to overcome this for experimental analysis and to train and test the model built, the dataset is split into training set and test set. 70% of the data is used for training the model and remaining 30% is used for testing the model.

In the first iteration of the preparation of the network data for further processing to build a clustering model for link prediction, it is processed to produce a similarity measure for further analysis. The similarity measure is calculated by extracting and processing the network node neighborhood features (here node degree would be used) using Jaccard Coefficient (JC), Adamic Adar (AA) and Preferential Attachment (PA). Thus, the computed identified similarity measure between the nodes of the network are arranged in ascending order and tabulated below in Table 4.7. The complexities of proposed method and existing local structural methods demonstrated in Table 4.8 [131].

Table 4.7 Extracted similarity measures for link prediction

Node *i*	**Node *j***	***JC***	**Node *i***	**Node*j***	***PA***	**Node *i***	**Node *j***	***Adamic Adar***
1412	3352	1	6	4	8845	6	4	19.95072059
3352	1412	1	4	6	8845	4	6	19.95072059
5254	1412	1	25	4	2610	25	4	16.34398298
5543	1412	1	5	25	2070	5	25	7.828409081
7478	1412	1	152	6	915	3	4	3.5801933
39	178	1	214	6	915	7	5	3.466104654
108	8283	1	75	6	915	30	54	1.481763337
152	214	1	7	25	810	54	30	1.481763337
178	39	1	3	4	725	55	30	1.468070074
182	271	1	282	6	610	28	30	1.459808553

214	152	1	645	6	610	38	30	1.245838453
271	182	1	650	6	610	8	30	1.012950461
286	182	1	38	25	360	299	30	1.012950461
300	182	1	1412	6	305	611	30	1.00468894
348	182	1	3352	6	305	61	30	0.948864248
349	182	1	5254	6	305	35	30	0.852166881
371	182	1	5543	6	305	75	30	0.838134815

The techniques applied over Wikipedia dataset are for extracting probability of having a legitimate link between the user *i* (node *i*) and user *j* (node *j*). The result obtained shows that Jaccard prediction among network node's using Jaccard coefficient is quite cumbersome. This is because the obtained similarity score of links for the node pairs is similar for most node pairs in the considered online social network. Due to this, appropriate clustering may also not possible. Figure 4.3 demonstrates the relative comparative analysis of similarity metrics applied over Wikipedia social network dataset. It is clearly observed that the similarity link prediction technique consistency varies with respect to the computation method.

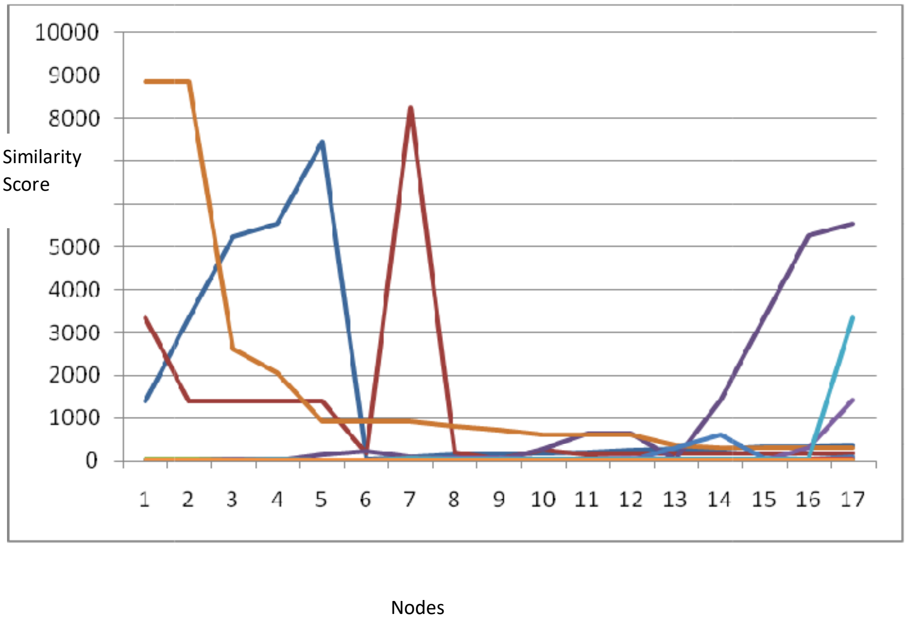


Figure 4.3 Sample representation of similarity measure distribution across node-pairs in the Network

The dataset used to train the model consists of similarity measure as an attribute obtained using three classical techniques namely preferential attachment, Adamic Adar and Jaccard Coefficient. In case *k*-means classification technique is applied with preferential attachment, the number of repetition are 5 with cluster sum of squared errors: 0.19278061143380593.Sum of squared errors is computed as the difference between the cluster mean and the observed value, it determines the changes in the cluster and for similar types of cases, the sum of squared error tends to be 0.

As the numbers of clusters are fixed to 2, *i.e.* the whole Wikipedia online social network will be divided in two dense communities that contain the potential node pairs with strong probability of having a future links between them. Although there are no missing values in our dataset but still can be treated by replacing them with the mean of the selected data. The final cluster centroids are represented in below table 4.8. The time taken to train and build the model (training data) is 0.02 seconds. As far as clustered instances are concerned cluster 1 consists of 1% of the nodes and cluster 2 of remaining 99%. Henceforth it has been observed that the probability of having a future links between the potential node pair is more in second cluster compared to first and will be dense as well. This shows that similarity between nodes in this cluster is high therefore prediction of link is also high.

In case of Jaccard Coefficient used as a similarity measure as an attribute the total number of repetitions for training the model was 2, within the cluster sum of squared errors 0.6845425386975178. The initial starting points randomly selected for cluster 1 and cluster are 1 and .209302 respectively. The final cluster centroids in the case of JC are .9941 for cluster 1 and .2297 for cluster 2.

Similarly, model is trained using *k*-means considering as similarity measure attribute has total number of iterations 2; that is, it has learnt in 2 iterations within the cluster sum of squared errors of 0.2649937282358271. The initial starting points randomly selected for the mean valued of cluster 1 and cluster 2 is 0.222232 and 0.1748816 respectively. Finally they obtained cluster centroids in the case of AA are 18.7485 for cluster 1 and .339 for cluster 2.

Table 4.8.The final clustering centroids

Attribute used (similarity measures)	Full Data	Cluster # 1	Cluster # 2
	343	335	8
Jaccard coefficient (JC)	0.9763	0.9941	0.2297
	343	3	340
Adamic Adar (AA)	0.5	18.7485	0.339
	343	2	341
Preferential Attachment (PA)	192.9446	8845	142.1994

Table 4.9 Clustered Instances of the online social networks

Clusters	Jaccard Coefficient*k*-Means		Adamic Adar*k*-Means		Preferential Attachment*k*-means	
1	8	2%	3	1%	2	1%
2	335	98%	340	99%	341	99%

It is visible from Table 4 and Table 5, the collaboration of *k*-means and similarity measures has performed effectively to group the network in two sub-networks, although the accuracy of the approach is still to be evaluated which is future scope of the research work. However, the density of nodes in the clusters-2 shows that the probability of prediction of links in between the nodes is higher than the other cluster (part of the network) which is also visible in Figure 4.6. (Figure 4.4 shows the clusters of similar nodes)

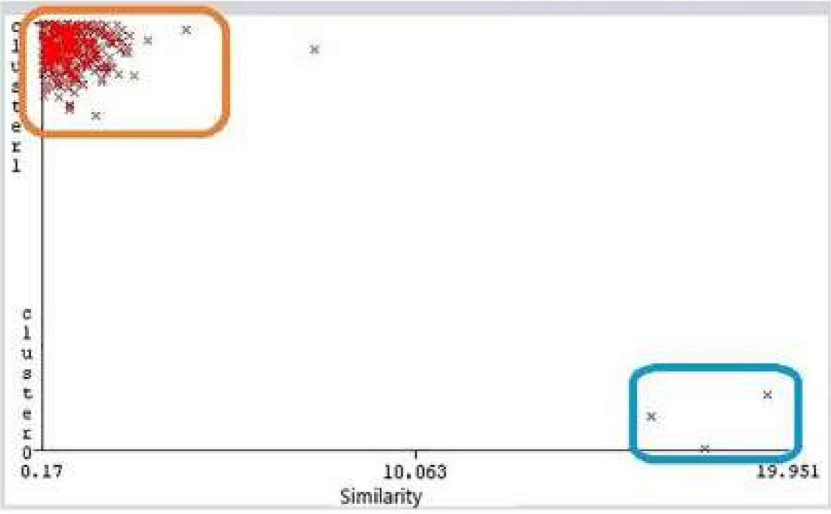


Figure4.4 Network cluster formed according to the similarity of the node pairs.

Figure 4.5 graphically demonstrates the obtained mean values of similarity measure attributes of respective node pairs for all the three classical link prediction methods. It has also been inferred from the comparative result anslysis shown in table that the model built using the features computed using Jaccard Coeeficient(JC), has wrongly classified 8 nodes leading false prediction of links between them, similarly the link prediction model with similarity feature computed using Adamic Adar has falsely classified 3 node pairs for link prediction. The model trained with features extracted with preferential attachment has an edge over these two and has identified 2 node pairs for predicting the future links.

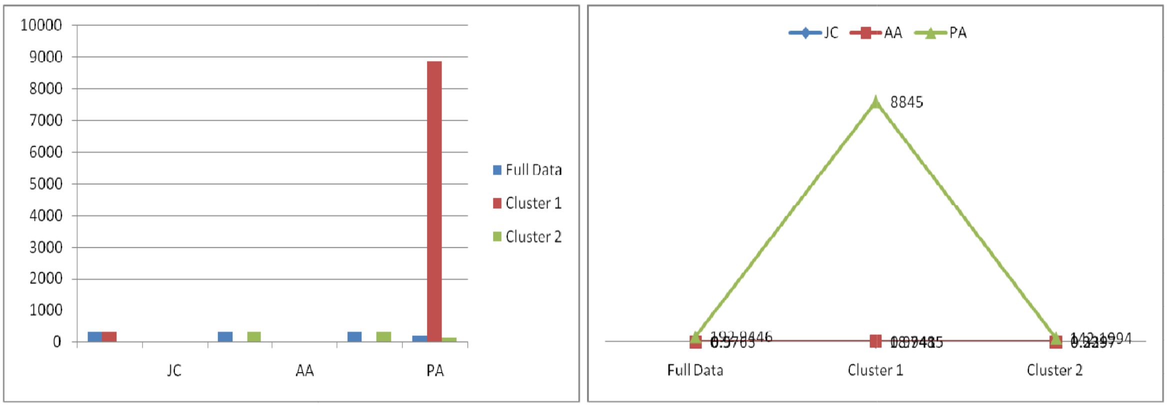
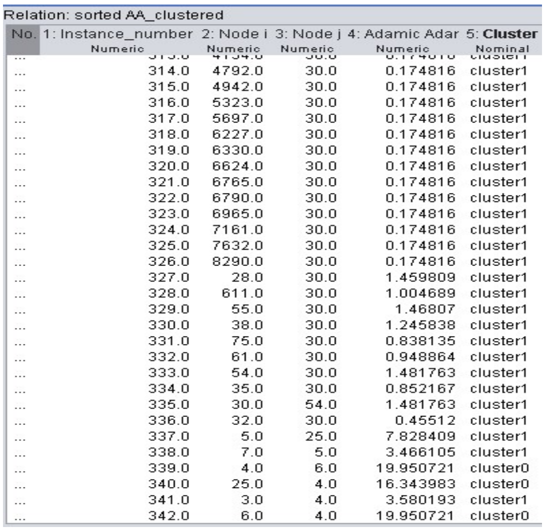


Figure 4.5 Similarity measure mean values as per which Wikipedia network grouping took place



Relation: sorted AA_clustered

No.	1: Instance_number Numeric	2: Node i Numeric	3: Node j Numeric	4: Adamic Adar Numeric	5: Cluster Nominal
...	313.0	[illegible]	30.0	0.174816	cluster1
...	314.0	4792.0	30.0	0.174816	cluster1
...	315.0	4942.0	30.0	0.174816	cluster1
...	316.0	5323.0	30.0	0.174816	cluster1
...	317.0	5697.0	30.0	0.174816	cluster1
...	318.0	6227.0	30.0	0.174816	cluster1
...	319.0	6330.0	30.0	0.174816	cluster1
...	320.0	6624.0	30.0	0.174816	cluster1
...	321.0	6765.0	30.0	0.174816	cluster1
...	322.0	6790.0	30.0	0.174816	cluster1
...	323.0	6965.0	30.0	0.174816	cluster1
...	324.0	7161.0	30.0	0.174816	cluster1
...	325.0	7632.0	30.0	0.174816	cluster1
...	326.0	8290.0	30.0	0.174816	cluster1
...	327.0	28.0	30.0	1.459809	cluster1
...	328.0	611.0	30.0	1.004689	cluster1
...	329.0	55.0	30.0	1.46807	cluster1
...	330.0	38.0	30.0	1.245838	cluster1
...	331.0	75.0	30.0	0.838135	cluster1
...	332.0	61.0	30.0	0.948864	cluster1
...	333.0	54.0	30.0	1.481763	cluster1
...	334.0	35.0	30.0	0.852167	cluster1
...	335.0	30.0	54.0	1.481763	cluster1
...	336.0	32.0	30.0	0.45512	cluster1
...	337.0	5.0	25.0	7.828409	cluster1
...	338.0	7.0	5.0	3.466105	cluster1
...	339.0	4.0	6.0	19.950721	cluster0
...	340.0	25.0	4.0	16.343983	cluster0
...	341.0	3.0	4.0	3.580193	cluster1
...	342.0	6.0	4.0	19.950721	cluster0

Figure4.6 Network cluster-0 and Network cluster-1 on the basis of similarity measure

The complexity performance of the algorithm with other state of the art local similarity-based techniques is demonstrated in Table 2. It shows that proposed algorithm takes lesser computation time and main memory while execution and superior with existing ones.

Illustratively the method has less computation cost and consumes less memory. The size of the large network is normalized and filtered to reduce it into relatively smaller data graph on the basis of feature scores. The feature score is extracted from neighbor nodes information. The proposed approach is superior in terms of time and space complexity compared to the existing methods. The complexity of the proposed model is reasonably less compared to the existing algorithms. The method proposed is effective in the case of dense social network. The method proposed is simple to implement and consumes relatively lesser time as it progresses for further iterations. The method proposed is highly scalable; it will perform effectively even when the online social networks scales up. Preciseness of link prediction is high compared to existing methods.

4.4.4 The Complexity Comparison

The complexity of the proposed method is evaluated with the complexities of existing state of the art structural based link prediction techniques. The proposed method is compared with JC, AA, and PA. The k –means shift supervised method used to converge the large online social network in to cluster having potential node pairs between which the probability of link existence is very high. Thus it reduces the number of iteration for further analysis of the subsequent obtained cluster network for link prediction. Therefore, compared to the complexity of of Jaccard Coefficient ($O(nk^2)$), Preferential Attachment ($O(n^2k^2)$) and Adamic Adar ($O(nk^2)$) link prediction techniques the proposed method is comparatively less complex and linear in nature.

4.5 Summary

The proposed link prediction approach developed for extraction of features which will contribute in prediction of future links between the potential node pair has been implemented and the obtained results are discussed. Online social network datasets have been used for training and testing of link prediction methods. It is observed that the proposed method is more efficient in identification of similarity between the nodes where existence of future link is to be predicted. The machine learning based supervised shift k means method generates cluster of nodes, and these nodes have higher probability of having connection in between.

Chapter 5

Conclusion and Future Works

5.1 Overview

In this research work the main aim was to understand the link prediction problem in online social network evolution and establishment of relationship between the users, to identify connection between users and predict new connection between them. It was also to devise a link prediction model which can efficiently predict links between the nodes in social networks. The procedure for performing the research involves many processes for achieving the set objectives of the research. The chapter here concludes the work along with subsequent outcomes of the research undertaken.

5.2 Conclusion

The goal of this work is to provide an idea to understand the online social network evolution by understanding the link prediction problem. Using the models and algorithms developed for link prediction several issues pertaining to real world communities can be addressed, whether it is any recommendation system for further decision making or investigation of criminal linkages in human networking. Therefore, eventually the thesis concludes following contributions.

1. Critically analyzed the evolution of undirected online social networks to understand the link prediction problem. In this process the existing models and techniques had been studied by applying classical local link prediction techniques (Jaccard, AA, PA and RA) on dynamic large social network namely; Wikipedia, Facebook, Deezer, GitHub and Twitch. The thorough study of link prediction gives us knowledge that links are representation of relations between the users of the social network and it can be of multiple types as people have relationships. It is observed that a link prediction technique provides similarity measures used to forecast future links between node pairs. Moreover the online social networks are dynamic in nature therefore the efficiency of techniques varies accordingly. Besides its nature the online networks are created for

specific purpose and community also such as coauthorship network, professional network etc.

2. Exploiting the node neighborhood attribute introduced a noval similarity algorithm that provides feature measure which will contribute in approximation of the future link between the node pairs. Since the computed measure contributes to the probability of the link prediction therefore named as probabilistic link prediction technique and it supersedes with the existing local link prediction techniques in terms of accuracy. While experimental analysis Jaccard was hundred percent accurate, but in the case of large complex network such accuracy is unrealistic and hence may be discarded, because it is a case of overfitting.
3. Use of similarity metrics for approximation of future association are the majorly accepted techniques in addressing the link prediction problem. The proposed probabilistic similarity measure considers the mutual neighbor node properties of node pair in a social network. As link prediction is a binary classification problem where possibility of a legitimate link between nodes is being obtained, therefore supervised machine learning framework is used to classify whether an interaction is there or not. The classifier will be learning the probability similarity feature extracted as to understand the significance of the probabilistic similarity measure. Standard online social network datasets have been used for the experimental analysis of the technique. It is evaluated using the standard performance parameter confusion metrics. The obtained results have shown that the proposed probabilistic similarity technique is significantly contributing in prediction of future links and has performed better. Moreover, the technique is performing better in evolution or online social network dynamically i.e. with respect to time. In future the amalgamation of similarity measures technique and machine learning techniques can be explored extensively for link prediction in social networks.

5.3 Future research opportunities

It is observed during the research work that mostly existing link prediction methods are using static network where nodes and links are fixed and will not going to change in future. But, real online social networks are dynamic in nature and evolving over the period of time. Though there are efforts in addressing the issue but still an effective, efficient and less complex method is yet to be evolved. For our study dynamic social networks are considered, although the method is exploiting structural attributes so can be applied on any type, but still other network dataset could be explored for further study of dynamic networks. The intense development in technology has raised cotemporary machine or bots pretending like humans, so there is equal possibility of having such proxy nodes in the network dataset, therefore such kinds of node identification will improve the validation of results.

Link prediction in the network from the perspective of future online social network evolution is studied, where the entire focus is on prediction of link which may create in near future. There are existing links between the nodes that are certainly inactive and likely to disappear in near future, may be due to less or no interaction between the users, so applying method for detection of such links on the basis of decreasing similarity could be a new research area to explore.

There are social networks wherein relationships between nodes can be of multiple types such as two persons can be professionally linked and may have informal and long friendship. Impact of such multiple relationships between the two nodes on future link prediction may be one area to explore. Link prediction in social networks such as Facebook, Github, Wikipedia and Twitch for link prediction is studied. Other types of large scale networks like biological networks for identification of protein interaction can be further explored for performance analysis of the proposed method.

It has been observed that lot of researches have been carried out for prediction of links which will likely to appear in near future. Prediction of relations which will likely to disappear in the future is one such interesting area that can give new understanding and findings about the further evolution of large scale social networks. Identification of potential nodes which can influence the its surrounding nodes and capable to change the overall behavior and structure of the social network may also substantially contribute in solving the link prediction problem along with other link attributes.

Bibliography

[1] Liben-Nowell D and Kleinberg J, “The link prediction problem for social networks,” 2003 Proceedings of the twelfth international conference on information and knowledge management, 3-8 November 2003, pp. 556-559.

[2] Michael Ley. Dblp.uni-trier.de: Computer science bibliography. http://dblp.uni-trier.de/xml

[3] M. E. J. Newman, “The structure and function of complex networks,” SIAM REVIEW, 45,2003.

[4] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical Structure and the Predictionof Missing Links in Networks,” Nature, 453, 2008.

[5] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins, “Microscopic evolution of social networks,” In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’08, pp. 462–470, 2008.

[6] Carpenter T, Karakostas G and Shallcross D, ”Practical Issues and Algorithms for, Analyzing Terrorist Networks,” Telcordia Technologies, viewed 12 June 2006, http://www.cas.mcmaster.ca/~gk/papers/wmc2002.pdf.

[7] Tang F, Mao C, Yu J and Chen J, “The implementation of information service based on social network systems,” Information Science and Service Science (NISS), 2011 5th International Conference on New Trends in, 2011. pp. 46 - 49.

[8] Travers J and Milgram S, “An Experimental Study of the Small World Problem” Sociometry, 1969, 32, 425.

[9] Tang F, Mao, C, Yu, J and Chen J, “The implementation of information service based on social network systems,” Information Science and Service Science (NISS), 2011 5th International Conference on New Trends in, 2011, pp. 46-49.

[10] H. Efstathiades, D. Antoniades, G. Pallis, M. D. Dikaiakos, Z. Szlávik and R. Sips, "Online social network evolution: Revisiting the Twitter graph," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 626-635, doi: 10.1109/BigData.2016.7840655.

[11] Barabasi A L, J H., Neda Z., Ravasz E, Schubert, A and Vicsek T, “Evolution of the social network of scientific collaborations,” Physica A: Statistical Mechanics and its Applications, 2002, 311, pp. 590-614.

[12] Akcora CG, Carminati B, and Ferrari E, “User similarities on social networks,” Social Network Analysis and Mining, 3, 2003, pp.475-495.

[13] Bartal A., Sasson E. and Ravid G, “Predicting Links in Social Networks Using Text Mining and SNA,” 2009 International Conference on Advances in Social Network Analysis and Mining, 2009, IEEE.

[14] Jelassi M N, Benyahia S and Mephu Nguifo E, “A personalized recommender system based on users' information in folksonomies,” 2013, Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion, ACM Press.

[15] Pavlov D, Mannila H and Smyth P, “Beyond independence: probabilistic models for query approximation on binary transaction data,” IEEE Transactions on Knowledge and Data Engineering, 15, 2003, pp. 1409-1421.

[16] Potts B B, Book Reviews: Networks: John Scott: Network Analysis: A Handbook. London, England, and New bury Park, CA: Sage Publications, 1992. Stanley Wasserman and Katherine Faust: Social Network Analysis: Methods and Applications, 1994.

[17] Lu L, Jin, C H and Zhou T, “Similarity index based on local paths for link prediction of complex networks,” 2009, Physical Review E, 80.

[18] L. A. Adamic and E. Adar, Social Networks 25, 211, 2003.doi:10.1016/S0378-8733(03)00009-1.

[19] Szwabe A, Ciesielczyk M, and Janasiewicz, T, “Semantically Enhanced Collaborative Filtering Based on RSVD,” Computational Collective Intelligence, Technologies and Applications, Springer Berlin Heidelberg, 2011.

[20] Kabbur, S, Ning, X and Karypis G, “FISM factored item similarity models for top-n recommender systems,” 2013 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13, ACM Press.

[21] Huang Z, Li X and Chen H, "Link prediction approach to collaborative filtering," 2005 Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05, ACM Press.

[22] Tang, F., Zhu, J., He, C., Fu, C., He, J. & Tang, Y. Scholat "An Innovative Academic Information Service Platform. Australasian Database Conference", 2016. Springer, 453-456.

[23] Tang F, Zhu J, Cao Y, Ma S, Chen Y, He J, Huang C, Zhao G and Tang Y, "PA Recommender: a pattern-based system for route recommendation," 2016 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016. AAAI Press, 4272-4273.

[24] Kimball Martin, "Graph Theory and Social Networks", Lecture Notes, April 30,2014.

[25] Deo N, "Graph Theory with Applications to Engineering and Computer Science," Prentice Hall, Upper-Saddle River, New Jersey,1974.

[26] McGlohon Mary, Akoglu Leman and Faloutsos Christos, "Statistical Properties of Social Networks," 10.1007/978-1-4419-8462-3_2, 2011.

[27] Sarr I., Missaoui R. "Temporal Analysis on Static and Dynamic Social Networks Topologies. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, 2011, https://doi.org/10.1007/978-1-4939-7131-2_387

[28] Nahla Mohamed Ahmed, Ling Chen, "An efficient algorithm for link prediction in temporal uncertain social networks", Information science, 331,pp102-136,2016.

[29] Pandey B., Bhanodia PK, Khamparia A., Pandey DK, "A comprehensive survey of edge prediction in social networks: Techniques, parameters and challenges," Expert Systems with Applications, Volume 124,2019,Pages 164-181,ISSN 0957-4174.https://doi.org/10.1016/j.eswa.2019.01.040.

[30] R. Zeng, Y. X. Ding and X. L. Xia, "Link prediction based on dynamic weighted social attribute network," 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, 2016, pp. 183-188.

[31] Yu K., Chu W., Yu S., Tresp V., and Xu, Z., “Stochastic relational models for discriminative link prediction,” Advances in Neural Information Processing Systems,19, pp.1553–1560, 2006.

[32] Murata, T., & Moriyasu, S. “Link prediction of social networks based on weighted proximity measures,” In IEEE/WIC/ACM International Conference on Web Intelligence, pp. 85–88, 2007. doi:10.1109/WI.2007.52.

[33] Brouard, C., d’Alché-Buc, F., & Szafranski, M., “Semi-supervised penalized output kernel regression for link prediction.”, In Lise Getoor, & Tobias Scheffer (Eds.),Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML’11) (pp. 593–600),2011, Omnipress.

[34] Lü, L., Jin, C. H., & Zhou, T. “Similarity index based on local paths for link prediction of complex networks”, Physical Review E, 80, 046122, 2009.

[35] G. H. Golub and C. F. Van Loan, “Matrix Computations,” Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.

[36] Jeh G and Widom J. “Simrank: A measure of structural-context similarity,” In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543, 2002.

[37] Song H. H., Cho T W, Dave V, Zhang, Y., and Qiu L. “Scalable proximity estimation and link prediction in online social networks,” In Proceedings of the 9th ACM SIGCOMM conference on Internet Measurement Conference, pp. 322–335, 2009.

[38] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” In Proceedings of the 19th International Conference on World Wide Web, pp. 641–650, 2010.

[39] R. N. Lichtenwalter, J. T. Lussier, and N.V. Chawla, “New perspectives and methods in link prediction,” In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining,pp. 243–252, 2010.
http://www.sciencedirect.com/science/article/pii/S138912860000044X

[40] J.L Herlocker, J.A. Konstann, K. Terveen, and J.T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Transaction on information system, 22(1),pp. 5-53,2004.

[41] Huang, Z., Li, X., and Chen, H., "Link prediction approach to collaborative filtering," in proceedings of the 5th ACM/IEEE-CS Joint conference on digital libraries ACM Press,2005.

[42] K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu, "Stochastic relational models for discriminative link prediction," Advances in Neural Information Processing Systems, vol-19, pp.1553, 2007.

[43] M. Bilgic, G.M. Namata, and L. Getoor, "Combining collective classification and link prediction," In Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 381–386,2007.

[44] K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu, "Stochastic relational models for discriminative link prediction," Advances in Neural Information Processing Systems, vol.-19, pp.1553–1560,2006.

[45] Z. Wang, J. Liang, R. Li, and Y. Qian, "An approach to cold-start link prediction: Establishing connections between non-topological and topological information," IEEE Transactions on Knowledge and Data Engineering, 28(11), 2857–2870, 2016. doi:10.1109/TKDE.2016.2597823.

[46] H. Wang, W. Hu, Z. Qiu, and B. Du, "Nodes' evolution diversity and link prediction in social networks," IEEE Transactions on Knowledge and Data Engineering, 29 (10), pp. 2263–2274, 2017. doi:10.1109/TKDE.2017.272852.

[47] B. Barzel, and A.L. Barabasi, "Network link prediction by global silencing of indirect correlations," Nature Biotechnology, vol-31, pp. 720–725, 2013.

[48] A. Narayanan, E. Shi, and B.I.P. Rubinstein, “Link prediction by deanonymization: How we won the Kaggle social network challenge,” In International Joint Conference on Neural Networks, pp. 1825–1834, 2013. doi:10.1109/IJCNN. 2011.6033446.

[49] Kunegis, J., and Lommatzsch, A, “Learning spectral graph transformations for link prediction,” In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 561–568). ACM New York. June 14 – 18, 2009.

[50] Dong, E., Jianping, L., Zheng, X., and Ning, W., “Bi-scale link prediction on networks. Chaos, Solitons and Fractals,” 78, pp.140–147, 2012.

[51] Wang, L., et al., “Link prediction by exploiting network formation games in exchangeable graphs,” In International Joint Conference on Neural Networks (IJCNN), pp. 619–626),2017. doi:10.1109/IJCNN.2017.7965910.

[52] Wu, J., Zhang, G., and Ren, Y., A balanced modularity maximization link prediction model in social networks. Information Processing & Management, 53(1), 295–307,2017. ISSN 0306-4573.

[53] Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M., “Link prediction using supervised learning,” In Proceedings of SDM Workshop of Edge Analysis, Counterterrorism and Security, 2006.

[54] Bai, L., Cui, L., Bai, X., and Hancock, E. R., “Deep depth-based representations of graphs through deep learning networks,” Neurocomputing in press, 2016. doi:10.1016/j. neucom.2018.03.087.

[55] Moradabadi, B., and Meybodi, M. R. A novel time series link prediction method: Learning autometa approach. Physica A: Statistical Mechanics and its Applications, 482, pp. 422–432, 2017.

[56] Shu, J., Chen, Q., Liu, L., and Xu, L., “A link prediction approach based on deep learning for opportunistic sensor network,” International Journal of Distributed Sensor Network, 13(4), pp. 1–7, 2017.

[57] Li, J.C., Zhao, D., Ge, B.-F., Yang, K.-W., and Chen, Y.-W., “A link prediction method for heterogeneous networks based on BP neural network,” Physica A, 495, 1–17,2018.

[58] Yin, L., Zheng, H., Bian, T., & Deng, Y., “An evidential link prediction method and link predictability based on Shannon entropy,” Physica A: Statistical Mechanics and its Applications, 482, pp. 699–712, 2017. https://doi.org/10.1016/j.physa.2017.04.106.

[59] Cui, W., Pu, C., Xu, Z., Cai, S., Yang, J., & Michaelson, A., “Bounded link prediction in very large networks,” Physica A, 457, pp. 202–214, 2016.

[60] Goa, M., Chen, L., Li, B., Liu, W., & Xu, Y.-c, “Projection-based link prediction in a bipartite network. Information Sciences,” 376, 158–171,2017.

[61] Herlocker, J. L., Konstann, J. A., Terveen, K., & Riedl, J. T., “Evaluating collaborative filtering recommender systems,” ACM Transactions on Information Systems,22(1),pp.5–53,2004.

[62] Lü, L. and Zhou T., “Link prediction in complex networks: A survey,” Physica A, 390, pp.1150–1170,2011.

[63] Y. Yang, R. N. Lichtenwalter, N.V. Chawla, “Evaluating link prediction methods,” knowl Inf Syst, 45,pp.751-782,springer-verlag, 2015.

[64] Sarukkai, R. R., “Link prediction and path analysis using Markov chains,” Computer Networks, 33(1-6), pp.377–386, 2000. https://doi.org/10.1016/S1389-1286(00) 00044-X.

[65] Wang, C., Satuluri, V., and Parthasarathy, S., “ Local Probabilistic models for link prediction,” Seventh IEEE International conference on Data Mining, ICDM 2007, 2007.

[66] Zhu, J. “Max-Margin Nonparametric Latent Feature Models for Link Prediction, Journal of latex class files, 6(1), 2007.

[67] Kashima H., Kato T., Yamanishi Y., Sugiyama M., and Tsuda K., “Link propagation: A fast semi-supervised learning algorithm for link prediction,” In Proceedings of International Conference on Data Mining, 2009. https://doi.org/10.1137/1.9781611972795.94

[68] Bliss, C. A., Frank, M. R., Danforth, M. R., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science, 5(5), 750–764.

[69] Kashima, H., & Abe, N. (2006). A parameterized probabilistic model of network evolution for supervised link prediction. In Proceedings of the Sixth International Conference on Data Mining (pp. 340–349).

[70] Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., & Elovici, Y. (2011). Link prediction in social networks using computationally efficient topological features. In IEEE Third International Conference on Social Computing (Socialcom) and IEEE Third International Conference on Privacy, Security, Risk and Trust (passat) (pp. 73–80).

[71] Lü, L., & Zhou, T. (2010). Link prediction in weighted networks: The role of weak ties. Europhysics Letters, 89, 18001.

[72] Liu, W., & Lü, L. (2010). Link prediction based on local random walk. Europhysics Letters, 89(5), 58007.

[73] Chiang, K.-Y., Natarajan, N., Tewari, A., & Dhillon, I. S. (2011). Exploiting longer cycles for link prediction in signed networks. In Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, & Ian Ruthven (Eds.), Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11) (pp. 1157–1162). ACM. https://doi.org/10.1145/2063576.2063742.

[74] Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. Journal of Systems and Software, 85(9), 2119–2132 2012.

[75] Liu, T., & Cerpa, A. E. (2011). Foresee (4C): Wireless link prediction using link features. In Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks (pp. 294–305).

[76] Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabási, A. L. (2011). Human mobility, social ties and link prediction. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1100–1108).

[77] Scellato, S., Noulas, A., & Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’11).

[78] Feng, X., Zhao, J. C., & Xu, K. (2012). Link prediction in complex networks: A clustering perspective. European Physical Journal B, 85(3). https://doi.org/10.1140/epjb/ e2011-20207-x

[79] Menon, A. K., & Elkan, C. (2011). Link prediction via matrix factorization. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.). Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science: vol 6912. Berlin, Heidelberg: Springer.

[80] Davis, D., Lichtenwalter, R., & Chawla, N. V. (2011). Multi-relational Link Prediction in Heterogeneous Information Networks. In International Conference on Advances in Social Networks Analysis and Mining (pp. 281–288). doi:10.1109/ASONAM.2011. 107.

[81] Sarkar, P., Chakrabarti, D., & Jordan, M. I. (2012). Nonparametric link prediction in dynamic networks. In Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML’12) (pp. 1897–1904).

[82] Yang, Z., Song, J., Huang, Z., Zhu, X., & Tian, H. (2014). A community-structure based adaptively optimized link prediction algorithm. In IEEE Fourth International Conference on Big Data and Cloud Computing (pp. 463–469). doi:10.1109/BDCloud. 2014.28

[83]Jiankun, Y., & Sili, F. (2014). A link prediction algorithm based on trust and similartag. In International Conference on Management of e-Commerce and e-Government (pp. 104–107). doi:10.1109/ICMeCG.2014.30.

[84] Guisheng, Y., Wansi, Y., & Yuxin, D. (2014). A new link prediction algorithm: Node link strength algorithm. In IEEE Symposium on Computer Applications and Communications (pp. 5–9). doi:10.1109/SCAC.2014.8.

[85] Fu, C. H., Chang, C. S., & Lee, D. S. (2014). Proximity measure for link prediction in social user-item networks. In Proceedings of IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014) (pp. 710–717). doi:10.1109/IRI.2014.7051959.

[86] Aouay, S., Jamoussi, S., & Gargouri, F. Feature based link prediction. In Proceedings of IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA) pp. 523–527, 2014. doi:10.1109/AICCSA.2014.7073243.

[87] Oczan, A., and Ögüdücü,S., G., "Multivariate temporal link prediction in evolving social networks", In proceedings of IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS) (pp. 185–190),2015. doi:10.1109/ICIS.2015.716659.

[88] Coskun, M., & Koyutürk, M., "Link prediction in large networks by comparing the global view of nodes in the network," In IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 485–492).2015.doi:10.1109/ICDMW.2015.195.

[89] Shalforoushan, S. H., & Jalali, M., "Link prediction in social networks using Bayesian networks." Mashhad, pp. 246–250.2015.doi: 10.1109/AISP.2015.7123483.

[90]Ahmed, C., & ElKorany, A., "Enhancing link prediction in Twitter using semantic user attributes," In The International Symposium on Artificial Intelligence and Signal Processing (AISP)n in Twitter using semantic user attributes, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1155–1161),2015. doi:10.1145/2808797.2810056.

[91]Malviya, V., & Gupta, G. P., "Performance evaluation of similarity-based link prediction schemes for social network," In 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 654–659),2015. doi:10.1109/NGCT.2015.7375202.

[92] Yang, Z., Hu, R., & Zhang, R., "An improved link prediction algorithm based on common neighbors index with community membership information," In 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)(pp. 90–93).2016. doi:10.1109/ICSESS.2016.7883022

[93] De, A., Bhattacharya, S., Sarkar, S., Ganguly, N., & Chakrabarti, S., "Discriminative link prediction using local, community, and global signals," IEEE Transactions on Knowledge and Data Engineering, 28(8), 2057–2070. doi:10.1109/TKDE.2016.2553665.

[94] Zhu, L., Guo, D., Yin, J., Steeg, G. V., & Galstyan, A., “Scalable temporal latent space inference for link prediction in dynamic social networks,” IEEE Transactions on Knowledge and Data Engineering,28(10),2765–2777,2016. doi:10.1109/TKDE.2016.2591009.

[95] Zhang, J., Chen, J., Zhi, S., Chang, Y., Yu, P. S., & Han, J. “Link prediction across aligned networks with sparse and low rank matrix estimation”. In IEEE 33rd International Conference on Data Engineering (ICDE),pp 971–982.2017.doi:10.1109/ICDE.2017.144.

[96] Wang, H., Hu, W., Qiu, Z., & Du, B., “Nodes evolution diversity and link prediction in social networks,” IEEE Transactions on Knowledge and Data Engineering, 29(10), 2263–2274. doi:10.1109/TKDE.2017.272852.

[97] Zhu, L., Guo, D., Yin, J., Steeg, G. V., & Galstyan, A., Scalable temporal latent space inference for link prediction in dynamic social networks (extended abstract). In IEEE 33rd International Conference on Data Engineering (ICDE),pp. 57–58, 2017. doi:10.1109/ICDE.2017.35.

[98] Wu, J., Zhang, G., & Ren, Y.,A balanced modularity maximization link prediction model in social networks, Information Processing & Management, 53(1), 295–307. ISSN 0306-4573. 2017.

[99] Zhang, Z., Wen, J., Sun, L., Deng, Q., Su, S., & Yao, P.,Efficient incremental dynamic link prediction algorithms in social network. Knowledge-Based Systems, 132, 226–235,2017.doi:10.1016/j.knosys.2017.06.035.

[100] Yin, L., Zheng, H., Bian, T., & Deng, Y.,An evidential link prediction method and link predictability based on Shannon entropy. Physica A: Statistical Mechanics and its Applications, 482, 699–712.2017. https://doi.org/10.1016/j.physa.2017.04.106

[101] Shang, K., Small M., & Yan W.S.,Link direction for link prediction. Physica A: Statistical Mechanics and its Applications, 469, 767–776. ISSN 0378-4371, 2017.

[102] Sharma, P. K., Rathore, S., & Park, J. H. Multilevel learning based modeling for link prediction and users consumption preference in online social networks. Future Generation Computer Systems,2017. https://doi.org/10.1016/j.future.2017.08.031.

[103] Yao, L., Wang, L., Pan, L., & Yao, K., Link prediction based on common-neighbors for dynamic social network. Procedia Computer Science,2016,83, 82–89.

[104] Srilatha, P., & Manjula, User behavior based link prediction in online social networks. In proceedings of International Conference on Inventive Computation Technologies (ICICT) (pp. 1–3). doi:10.1109/INVENTIVE.2016.7823266.

[105] Ströde, V., Campos, F., Pereira, C. K., Zimbrão, G., & Souza, J. M.,Information Extraction to improve link prediction in scientific social networks. In IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 515–520),2016. doi:10.1109/CSCWD.2016.7566043

[106] Gupta, A. K., & Sardana, N., "Naïve Bayes approach for predicting missing links in ego networks", In IEEE International Symposium on Nanoelectronic and Information Systems (iNIS) (pp. 161–165). 2016.

[107]Wang, Y., & Bai, L.,Link prediction via supervised dynamic network formation. In 23rd International Conference on Pattern Recognition (ICPR) (pp. 4160–4165),2016. doi:10.1109/ICPR.2016.7900286.

[108] Hours, H. Fleury, E., and Karsai, M., Link prediction in the twitter mentioned network:impactss of local structure and similarity interest. in IEEE 16th international conference on Data mining workshop (ICDMW), pp 454-461. 2016.

[109] Laishram, R., Mehrotra, K., & Mohan, C. K.,Link prediction in social networks with edge aging. In IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI),pp.606–613,2016.

[110] Amin, M. I., & Murase, K.,Link prediction in scientists collaboration with author name and affiliation. In Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS),pp.233–238,2016.

[111] Shu, J., Chen, Q., Liu, L., & Xu, L.A link prediction approach based on deep learning for opportunistic sensor network. International Journal of Distributed Sensor Network,13(4), 1–7, 2017.

[112] Yan, B., & Gregory, S., Finding missing edges in networks based on their community structure. arXiv:1109.2793,2012.

[113] Li, J.C., Zhao, D., Ge, B.F., Yang, K.W., & Chen, Y.W., "A link prediction method for heterogeneous networks based on BP neural network," Physica A,495,pp1–17,2018.

[114] Li, Y., Luo, P., Fan, Z.-p., Chen, K., & Liu, J., "A utility-based link prediction method in social networks", European Journal of Operational Research, 260(2), 693–705,2017.

[115] Shakibian, H., Charkari, H. M., & Jalili, S., "A multi layered approach of link prediction in heterogeneous complex network," Journal of Computational Science, 17, 73–82,2016.

[116]Ahmed, N. M., & Chen, L., "An efficient algorithm for link prediction in temporal uncertain social networks," Information Science, 331, 102–136, 2016.

[117] Aghabozorgi, F., & Khayyambash, M. R., "A new similarity measure for link prediction based on local structures in social networks," Physica A, 501, 12–23,2018.https://doi.org/10.1016/j.physa.2018.02.010.

[118] Yasami, Y., & Sagaei, F., "A novel multi layer model for missing link prediction and future link forecasting in dynamic complex networks", Physica A: Statistical Mechanics and its Application, 492, 2166–2197,2018.

[119] Mohan, A., Venkatesan, R., & Pramod, K. V., "A scalable method for link prediction in large real world networks", Journal of Parallel and Distributed Computing, 109, 89–101,2017.

[120] Dong, E., Jianping, L., Zheng, X., & Ning, W., "Bi-scale link prediction on networks. Chaos," Solitons and Fractals, 78, 140–147,2015.

[121]Bo, Z., Huan, Z., Meizi, L., Quin, Z., & Jifeng, H., "Trust Traversal: A trust link detection scheme in social network", Computer Networks, 120, 105–125,2017.

[122] Nguyen-Thi, A.-T., Nguyen, P. Q., Ngo, T. D., & Nguyen-Hoang, T. A. "Transfer AdaBoost SVM for link prediction in newly signed social networks using explicit and PNR features", Procedia Computer Science, 60, 332–341,2015.

[123] Sherkat, E., Rahgozar, M., & Asadpour, M., “Structural link prediction based on ant colony approach in social networks,” Physica A, 419, 80–94, 2015.

[124]Shakibian, H., & Charakri, N. M., “Statistical similarity measures for link prediction in heterogeneous complex networks,” Physica A: Statistical Mechanics and its Applictions, 501(1), 248–263,2018.

[125] Goa, M., Chen, L., Li, B., Liu, W., & Xu, Y., “Projection-based link prediction in a bipartite network,” Information Sciences, 376, 158–171,2016.

[126] He Y.L., Liu J.N.K., Hu Y. X., and Wang X.Z., “OWA operator based link predeiction ensemble for social network,” Expert System with Application, 42,21-50,2015.

[127] Lin D. “An information-theoretic definition of similarity,” Paper presented at: Proceedings of the 15th International Conference on Machine Learning; 1998:296-304.

[128] Leicht EA, Holme P, Newman MEJ. “Vertex similarity in networks,” Phys. Rev. E. 2006;E73.73(2):026120. https://doi.org/10.1103/PhysRevE.73.026120.

[129] Koren Y, Bell RM, Volinsky C. “Matrix factorization techniques for recommender system,” IEEE Computer, 2009; 42(8):30-37.

[130]Singh AP, Gordon GJ. “A unified view of matrix factorization models.” Paper presented at: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases Part II. Number 16 in ECML PKDD; 2008:358–373; Springer Verlag.

[131] Agrawal P, Garg VK, R. Narayanam. “Link label prediction in signed social networks.” Paper presented at: Proceedings of the 23rd International Joint Conference on Artificial Intelligence; 2013.

[132] Ye, J, Cheng H, Zhu Z, Chen M. “Predicting positive and negative links in signed social networks by transfer learning,” Paper presented at: Proceedings of the 22nd International Conference on World Wide Web;2013:1477–1488.

[133] Candes EJ, Plan Y. “Matrix completion with noise,” Proceedings IEEE. 2010;98(6):925-936. https://doi.org/10.1109/ JPROC.2009.2035722.

[134] Wang, YX, Xu H. “Stability of matrix factorization for collaborative filtering,” Paper presented at: Proceedings of the 29th International Conference on International Conference on Machine Learning; 2012 :163–170.

[135] Zhou Z, Li X, Wright J, Candes EJ, Ma Y. “Stable principle component pursuit.”, Paper presented at: Proceedings of the IEEE International Symposium on Information Theory; 2010:1518–1522; Austin.TX.

[136] Chen P. “Optimization algorithms on subspaces: re-visiting missing data problem in low-rank matrix”, Int. J. Comput, Vis. 2008; 80(1):pp125-142.

[137] Ozcan A, Oguducu SG. “Link prediction in evolving heterogeneous networks using the NARX neural networks,” Knowl Inf Syst. 2017;55920:pp333-360.

[138] Mallek S, Boukhris I, Elouedi Z. Ericlefgraveevre, “Evidential link prediction in social networks based on structural and social information,” J Comput Sci. 2018;30:98-107. https://doi.org/10.1016/j.jocs.2018.11.009.

[139] Faloutsos, M, Faloutsos P, and Faloutsos C, “On power-law relationships of the Internet topology,” In Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication (SIGCOMM '99). ACM, New York, NY, USA, 1999 pp.251-262. DOI: https://doi.org/10.1145/316188.316229.

[140] Kumar R, Raghavan P, Rajagopalan S and Tomkins A, “Trawling the Web for emerging cyber-communities,” In Proceedings of the eighth international conference on World Wide Web (WWW '99), Philip H. Enslow, Jr. (Ed.). Elsevier North-Holland, Inc., New York, NY, USA,1999, pp. 1481-1493

[141] Barabasi A, and Albert R, “Emergance of scaling in random networks,” Science, 1999, pp. 286:509-512

[142] Kumar R, Novak, J, Raghavan, P, and Tomkins, A, “Structure and evolution of blogspace. Commun” ACM 47, 12 December 2004, pp. 35-39.doi: https://doi.org/10.1145/1035134.103516.

[143] Sapountzi A and Psannis KE, “Social networking data analysis tools and challenges,” Future Generation Computer Systems, 2018, Vol-86, pp.893-913, ISSN 0167-739X. https://doi.org/10.1016/j.future.2016.10.019

[144] Peng S, Zhou Y, Cao L, Yu S, Niu J. and Jia W. “Influence analysis in social networks: A survey,” Journal of Network and Computer Applications,2018, Vol-106,pp.17-32, ISSN 1084-8045.https://doi.org/10.1016/j.jnca.2018.01.005.

[145] Leppink,J.&Fuster, P. (2018).Social Networks as an Approach to Systematic review. Health Professions Education.ISSN 2452-3011.https://doi.org/10.1016/j.hpe.2018.09.002

[146] Fetterly M, Manasse M, Najork and Wiener J, “A large-scale study of the evolution of web pages,” Software Practice and Experience, 2004, 34(2), pp. 213–237.

[147] Ntoulas Cho J, and Olston C, “What’s new on the web? The evolution of the web from a search engine perspective,” In 13th WWW.pp. 1–12

[148] http://igraph.org/python/doc/tutorial/tutorial.html

[149] Bhanodia PK, Khamparia A., Pandey B., Prajapat S., “Online Social Network Analysis: Hidden Link Prediction in Stochastic Social Networks,” 2019,Pages:14, DOI: 10.4018/978-1-5225-9096-5.ch003, IGI Global.

[150] J. McCauley and J. Leskovec, “Learning to Discover Social Circles in Ego Networks,” NIPS, 2012.

[151] Benedek Rozemberczki and Carl Allen and Rik Sarkar,"Multi-scale Attributed Node Embedding", arXiv, 2019.

[152] Benedek Rozemberczki and Rik Sarkar,"Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models", CIKM 2020. https://arxiv.org/abs/2005.07959

[153] Witten, I & Frank, E 2005, “Data Mining: Practical Machine Learning Tools and Techniques,” second edition, Morgan Kauffman.

[154] Carletta, J 1996, “Assessing agreement on classification tasks: the kappa statistic,” Computational Linguistics, vol. 22, no. 2, pp 249-254, viewed 12 June 2006, http://acl.ldc.upenn.edu/J/J96/J96-2004.pdf.

[155] Fleiss JL, “Statistical methods for rates and proportions,” second edition, John Wiley & Sons, New York, 1981.

[156] Kaufman R, Leonardo PJ. “Finding Groups in Data: An Introduction to Cluster Analysis,” Hoboken, NJ: JohnWiley & Sons; 2009:344

[157] Burkardt T. K-means clustering. Virginia Tech. “Advanced research computing. Interdiscipline Center Appl Math. 2009.

[158] https://snap.stanford.edu/data

List of Research Publications

[1] Bhanodia PK, Khamparia A, Pandey B. Supervised shiftk-means based machine learning approach for link prediction using inherent structuralproperties of large online social network. Computational Intelligence. 2020;1–18.https://doi.org/10.1111/coin.12372. **(SCI; Imapct Factor 5.452)**

[2] Pandey B., Bhanodia PK, Khamparia A., Pandey DK,A comprehensive survey of edge prediction in social networks: Techniques, parameters and challenges,Expert Systems with Applications,Volume 124,2019,Pages 164-181,ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2019.01.040. **(SCI; Impact Factor 1.196)**

[3]Bhanodia PK., Khamparia A., Pandey B., An Approach to Predict Potential Edges in Online Social Networks. In: Jat D., Shukla S., Unal A., Mishra D. (eds) Data Science and Security. Lecture Notes in Networks and Systems, vol 132. Springer, Singapore.2021.http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-981-15-5309-7_1. (**Scopus**)

[4]Bhanodia PK., Khamparia A., Pandey B., An efficient link prediction model using supervised learning. Internaltonal conference on intelligent communication and computational research (ICICCR-2020), January 2020, Springer nature: In Studies com. Intelligence vol.921, Ashish Khanna et. Al. (Eds): Recent Studies on computational intelligence, Springer Nature. **(scopus)**

[5]Praveen Kumar Bhanodia, Aditya Khamparia,Babita Pandey,and Shaligram Prajapat,Online Social Network Analysis: Hidden Link Prediction in Stochastic Social Networks, 2019,Pages:14, DOI: 10.4018/978-1-5225-9096-5.ch003, IGI Global.

[6]Praveen Kumar Bhanodia,Kamal Kumar Sethi, Aditya Khamparia, Babita Pandey, Shaligram Prajapat,Similarity-Based Indices or Metrics for Link Prediction, Hidden Link Prediction in Stochastic Social Networks, 2019, pages-29. DOI: 10.4018/978-1-5225-9096-5.ch00