

**Prediction of New Citation Links in
Author-Author Directed Network**

Mujtaba JAWED
Master Thesis

Department: Computer Engineering
Supervisor: Prof. Dr. Mehmet KAYA

March-2016

**REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

**PREDICTION OF NEW CITATION LINKS IN AUTHOR-AUTHOR
DIRECTED NETWORK**

**Master Thesis
Mujtaba JAWED
(121129120)**

Submission Date to the Institute: 23 February 2016

Thesis Presentation Date: 11 March 2016

**Advisory Committee: Prof. Dr. Mehmet KAYA (Supervisor)
Prof. Dr. Ali KARCI
Asst. Prof. Dr. Taner TUNCER**

March, 2016

ACKNOWLEDGMENT

Thanks God before everything and after everything for giving me the knowledge and ability to complete this research work in this final form.

I am very thankful and indebted to my respected teacher Prof. Dr. Mehmet KAYA for his valuable and thorough supervision throughout all the phases of my thesis, without which it wouldn't be possible to be completed. Not just during my thesis preparation but from the time when I came to Turkey he is one of the persons who helped me very much.

My special thanks goes to Prof. Dr. Ali KAECI and Asst. Prof. Dr. Taner TUNCER for taking part as an advisory committee in my thesis presentation and their inestimable feedbacks which enhanced and improved my research.

I would like to express my gratitude and special thanks to Turkey Government and Presidency for Turks abroad and related communities (YTB (Yurtdışı Türkler ve Akraba Topluluklar Başkanlığı)) for providing the master degree scholarship to me, by which I found the ability to become familiar to Turkish people, Turkish culture. Their unlimited helps, supports and encouragements are greatly appreciated.

Acknowledging my beloved family for their supports and encouragements in the hard times, I am forever indebted to my family especially my mother and my father for all their helps both materially and morally. I would like to thank my mother and my aunt for constantly including me in their prayers.

My special thanks goes to my dear friends and all faculty members for their helps.

TABLE OF CONTENTS

Acknowledgment.....	I
Table of Contents.....	II
List of Figures.....	IV
List of Tables	V
Abstract.....	VI
1. Introduction.....	1
1.1 Objectives.....	2
1.2 Organization of the thesis.....	3
2. Social Networks	4
2.1 What is a network?.....	4
2.2 Social Networks	4
2.3 Definition	5
2.4 Types of social networks.....	7
2.4.1 Relational-based types of social networks	7
2.4.2 Directional-based types of social networks.....	8
2.5 Social Network Analysis.....	9
2.5.1 Basic Metrics in Social Network Analysis.....	11
3. Link Prediction.....	16
3.1 Unsupervised Approach.....	17
3.2 Supervised Approach	17
3.3 Related Works.....	18
3.4 Proximity Measures in Link Prediction	19
3.4.1 Number of Common Neighbors (CN).....	19
3.4.2 Jacquard's coefficient.....	20
3.4.3 Preferential Attachment (PA).....	21
3.4.4 Adamic-Adar Coefficient (AA).....	21
3.4.5 Path Distance (PD)	22
3.4.6 Resource Allocation (RA)	22
3.4.7 Local Path (LP).....	23
4. Link prediction in directed citation network.....	25
4.1 Directed Networks	25

4.2	Background	26
4.3	Link Prediction.....	27
4.4	Used Traditional Proximity Measures	28
4.4.1	Number of Common Neighbors (CN).....	28
4.4.2	Jacquard's coefficient.....	28
4.4.3	Preferential Attachment (PA).....	29
4.4.4	Adamic-Adar Coefficient (AA).....	29
4.5	Citation Network and Tried Patterns.....	30
4.6	The Proposed Time-Frame Based Link Prediction in Directed Networks	32
4.6.1	Temporal Structure.....	32
4.6.2	Temporal event.....	32
4.7	Time Frame Based Score	35
4.8	Experimental Results	36
5.	Link Prediction in Weighted Citation Networks.....	38
5.1	Weighted Networks.....	38
5.2	Measures for Weighted Networks.....	39
5.3	Link Prediction.....	39
5.4	Traditional proximity measures	41
5.4.1	Number of Common Neighbors (CN).....	41
5.4.2	Jacquard's coefficient.....	41
5.4.3	Preferential Attachment (PA).....	42
5.4.4	Adamic-Adar Coefficient (AA).....	42
5.5	Citation network and triad patterns	43
5.6	The Proposed Time-Frame Based Link Prediction in Directed Networks	46
5.6.1	Temporal Structure.....	46
5.6.2	Temporal event.....	46
5.6.3	Weighted temporal event.....	47
5.7	Time Frame-based Score.....	49
5.8	Experimental Results	51
6.	Conclusion.....	53
	References.....	55
	Curriculum vita	59

LIST OF FIGURES

Figure 2.1 A simple Social Network_____	5
Figure 2.2 Social Networks over Internet _____	6
Figure 2.3 Average hours spent on social networking site per visitor across Europe _____	6
Figure 2.4 A directed Social Network with 10 vertices (or nodes) and 13 edges. _____	8
Figure 2.5 An undirected Social Network with 10 and 11 edges. _____	9
Figure 2.6 An example network for degree centrality calculation _____	11
Figure 2.7 An example network for closeness centrality calculation _____	12
Figure 4.1 Triad graph patterns _____	30
Figure 4.2 Closed F1X triads _____	31
Figure 4.3 Closed F2X triads _____	31
Figure 4.4 Closed F3X triads _____	31
Figure 4.5 Protective, inventive and regressive events _____	34
Figure 5.1 A simple weighted network _____	38
Figure 5.2 Citation network _____	43
Figure 5.3 Triad graph patterns _____	44
Figure 5.4 Closed F1X triads _____	45
Figure 5.5 Closed F2X triads _____	45
Figure 5.6 Closed F3X triads _____	45
Figure 5.7 Protective, inventive and regressive events _____	49

LIST OF TABLES

Table 1 Precision values (%) of five methods	37
Table 2 Precision values (%) of five methods on the weighted network	51
Table 3 AUC values of the proposed method for different values of α	52

ABSTRACT

Link prediction in weighted and directed networks according to the application of temporal information can be point as an important problem in social network analysis. Link prediction tends to guess the likelihood of the connections occurrence between nodes. In addition the link prediction aims to determine the missing links in the network, which uses the state of the network up to a given time for predicting the new links in future. Most of the previous works have deployed to unweighted or undirected networks and for computing the proximity scores, only the current state of the network has considered without taking any temporal information into account, which can be point as a limitation in link prediction studies. In this study we tried to overcome the above mentioned limitation by analyzing the development of topological measures in a weighted-directed citation network on a specific period of time. For achieving this aim, chosen similarity metric deployed to all non-connected pairs of nodes in different frames of time in the network. Then, time frames are built for each pair to record their values which provided by the metric. Experiments on unsupervised prediction on a weighted-directed citation network show that the proposed method finds satisfactory results and is promising.

Keywords: Social networks, Social network analysis, Link prediction, directed networks, weighted networks, citation networks, temporal events.

1. INTRODUCTION

Humans are creatures of social and they interact with each other in several manners, like these human social networks are everywhere present in the nature. Such social networks have wide area of application which can be range from offline networks based on kinship ties, friendship to online networks. For instance citation network is an example of offline networks and in online social networks Face-book, Myspace, LinkedIn, etc.... are the popular once. The social networks represent as graphs or hyper graphs and there is an extensive body of literature on social networks [1].

The advent and development of social networks is one of the most exciting events in recent years. The continuously developments of social networks generated a big source of information in different fields which has been attracting attention of many researchers. In literature of social networks the researches have proposed so many kinds of predictive problems. Link prediction is one of the mentioned problems which predict the links that tend to appear between a pair of nodes in a network in the future. Liben-Nowell and Kleinberg [2] proposed the link prediction problem for the first time, which has been studying widely from that time and continuing till now. A link between two individuals is established base on many elements. Existing links give the individuals more chances to meet new people who can become their friends in future.

While the link prediction problem has been applied in many domains like social networks, protein-protein inter-action, record linkage problem, web-link prediction etc., we restrict our study to application of link prediction in social networks especially prediction of new citation links in Author-Author directed network. The link prediction problem has a wide range of applications in addition to the well-known example of e.g., recommendation systems [3] [4], making recommendation to create mutually beneficial professional links [5], improve navigational efficiency of websites [5] etc.

In this study already existed citation network of authors will be considered according to the different farms of time from the past till a specific period of time near to the present. A Dataset that include citation relationship of authors will be considered. By using of this dataset the citation network between authors will be designed. In first instance we will obtain already existed links of the network on a time farm based like year to year or each two years. According

to the number of citation of authors a weight will be assigned to the links in the network therefore the mentioned network is going to change to a directed and weighted network. Within the scope of this study a real dataset which will be used is going to abstract from DBLP and CiteseerX websites. By using of obtained links attributes and applying the proximity measures prediction of new links between those nodes that don't have any relation with each other will be calculated.

1.1 Objectives

In most of the link prediction studies the topological attributes, static attributes or features of vertices and links have been considered for finding or predicting the upcoming/future connections. Temporal information of vertices and connections has been taken in account in some of studies. Even though most of the static features provide valuable information about general social phenomenon that can be used for predicting the powerful connections, temporal attributes have a big effect on link evolution. It is not true all the time that the static features don't change over the time. It is worthy to study how to use the information gained from temporal behavior of vertices and connections to predict future powerful connections. Therefore we set the (time frame based link prediction in directed-weighted citation networks) as a main goal of this thesis.

Different from previous studies in this thesis we applied four topological proximity metrics (Common Neighbor index, Jacquard's Coefficient index, Preferential Attachment index and Adamic Adar index) both weighted and unweighted versions to the link prediction problem in a real citation network which is directed and weighted. Also the application of temporal information has considered in this study.

In the preformed experiments, we did a comparison between weighted and unweighted proximity metrics, according to the experimental results we can highlight that weighted proximity metrics outperform the unweighted metrics.

Experimental results applied on real dataset shows that the proposed method gives accurate prediction and promising results which outperform the traditional proximity metrics.

As we mentioned above all of the preformed experiments are according to the common neighbor based metrics and we didn't consider the global information based metrics and path based metrics, therefore the application of mentioned metrics can be consider as future works,

moreover this work still has limits regarding the case study, which the experiments deployed on one dataset and as future studies this work can extend by applying on different datasets.

1.2 Organization of the thesis

The rest of this thesis is organized as follow:

Chapter 2: Social network brings a short definition of network, information about social networks, types of social networks and social networks analysis.

Chapter 3: Link prediction starts with the two well-known approaches of link prediction. The past researches related to the link prediction are addressed in this chapter and the proximity measures in link prediction are described at end.

Chapter 4: Link prediction in directed citation network introduce a novel method for predicting new links in directed citation network based on temporal information. The effectiveness of temporal information in link prediction is shown by the experimental results at the end of this chapter.

Chapter 5: Link Prediction in Weighted Citation Networks starts with a short information about weighted networks. A novel method for predicting new links in weighted-directed citation network based on temporal information is addressed in this chapter.

Chapter 6: Conclusion summarizes and discusses the contributions of this research work. It also discusses the limitations of the features introduced in this research work, and future directions of our research.

2. SOCIAL NETWORKS

2.1 What is a network?

Collection of relations is called network. Formally a network consist an object's set which mathematical calls nodes and a set of relations among objects or nodes. At least a simple network consist two nodes like the 1 and 2 in the below that might be two person or two authors, and a connection that connect them which can be “working in the same office or citing each other”



Relations can have direction in a network, as an example in the below 1 cite the 2 while there is no citation relation from 2 to 1.



Also relations in a network can be bidirectional or mutual like the one below in which both nodes (authors) cited each other mutually.



When there is more than one kind of relations in a network then it is called multiplex relations. For instance if 1 and 2 work in the same university and they cite each other then they share a multiplex relation among them.

2.2 Social Networks

Social networks have been increasing largely during the last 100 years so lots of people are interested to take part in social networks like Myspace Facebook, YouTube, Twitter, etc. to make connection with their families, friends, colleaguesetc. in a virtual environment.

2.3 Definition

A graph $G = (V, E)$ including a set of vertices V and a set of ties/edges E , which the vertices show objects of the graph like people, groups, or foundations and the ties show the relations between objects of the network like friendship, family relations, work relations, etc.

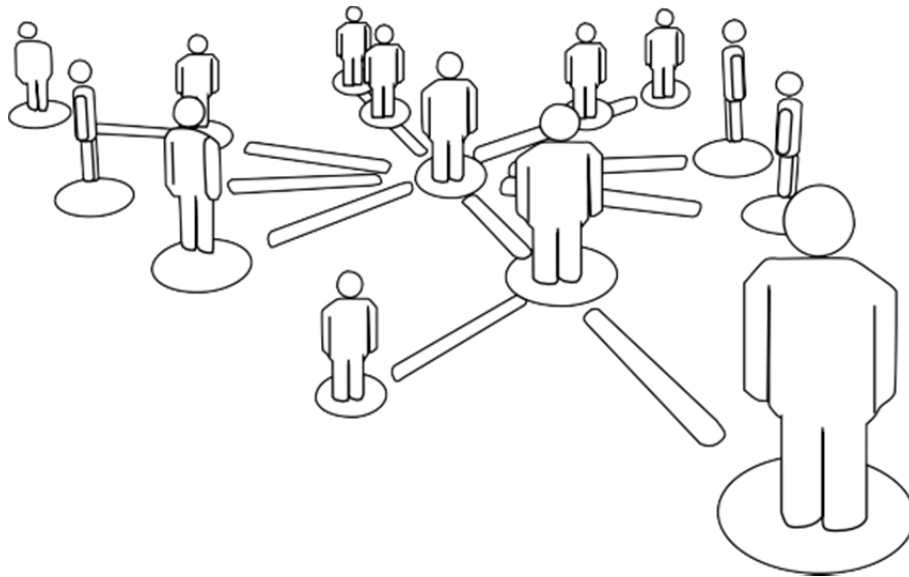


Figure 2.1 A simple Social Network

As mentioned above social networks can be shown by persons or foundations and the relationships of them. Generally these relationships show one or different kinds of solidarity (like religion and thoughts) or more special relations (such as citation relations and kinship).

The cardinality of nodes expands dynamically, especially on the Web, as new documents and profiles are created continuously while further populating additional nodes in the social Web [7].

Technology and internet developments made it easy and so simple to grant accessing of social networking sites therefore use of social networking sites become a habit to a majority of people over the world. Facebook, Twitter, Google+, LinkedIn, YouTube, Tencent QQ... are the most popular social networking sites which are using around the world.



Figure 2.2 Social Networks over Internet

Social networks are growing day by day, though people desire to spend time in these networks because people desire to make communities, to extend their communities, to be involved in more cases and activates, people want to connect and reconnect. According to a research which had been done about (average hours spent on social networking site per visitor across Europe) by comScore Data Mine in 2010 shows that “Spanish Internet users between the age of 15-24 spent most time on social networking sites with an average of 11 hours in December 2010, followed by 15-24 year olds from the UK and Italy. 35-54 year old Internet

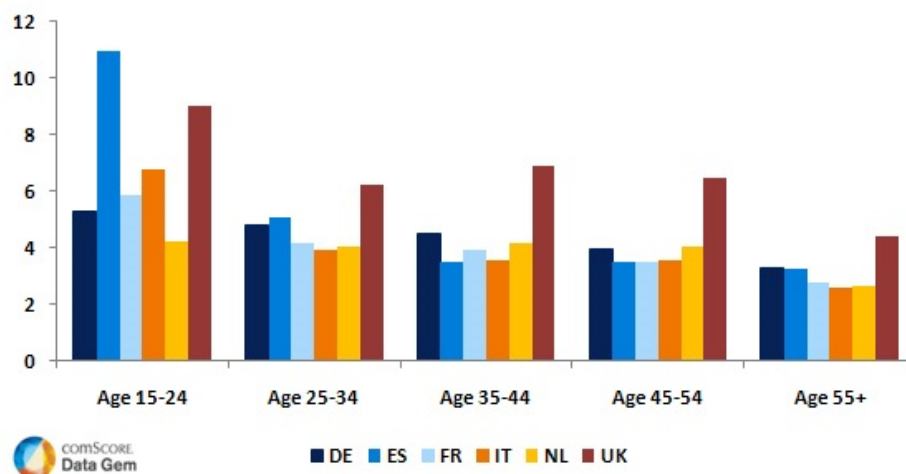


Figure 2.3 Average hours spent on social networking site per visitor across Europe

users in the UK spent on average more time on Social Networking sites than their 25-34 year old counterparts.”[8]

2.4 Types of social networks

There are different types of social network in literature of social network concept which have considered by researches according their usages. Forasmuch as the topic that will be studied by us (Prediction of New Citation Links in Author-Author Directed Network) is related to directed and multi relational social networks therefore relational-based and directional-based types of social networks will be considered shortly.

2.4.1 Relational-based types of social networks

There are two types of Relational-based social networks homogeneous and heterogeneous. [8]

2.4.1.1 Homogeneous

In homogeneous social networks, objects have just one kind of relation and the information interchange between objects take place through this relation. Thus, social network analysis considers only one type of knowledge exchange between network elements.

2.4.1.2 Heterogeneous

A multi-relational social network which represents different kinds of relationships among objects is a Heterogeneous social network. In this kind of social networks the exchange of knowledge between objects takes place through different kinds of relationships. Thus, multi-relational social network analysis assumes that elements are exchanging different types of knowledge depending on the types of relationships linking them.

2.4.2 Directional-based types of social networks

2.4.2.1 Directed social networks

Actually a directed network or a digraph is a directed graph. A directed network is a graph or a set of objects (called vertices or nodes) that are connected together, where all the edges are directed from one vertex to another. In contrast, a graph where the edges point in a direction is called a directed social network. When drawing a directed graph, the edges are typically drawn as arrows indicating the direction, as illustrated in the following figure.

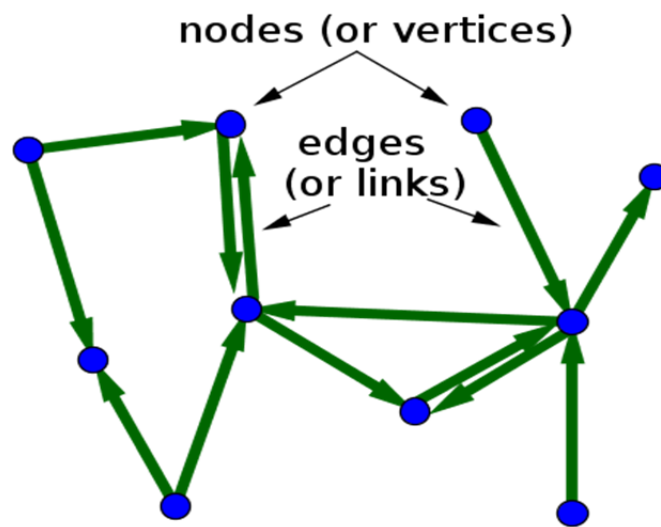


Figure 2.4 A directed Social Network with 10 vertices (or nodes) and 13 edges.

Formally a directed Social Network/Graph can be defined as $G = (N, E)$, consisting of the set N of nodes and the set E of edges, which are ordered pairs of elements of N .

2.4.2.2 Undirected Social Networks

An undirected social network is actually an undirected graph. An undirected network is a graph or a set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional. When drawing an undirected graph, the edges are typically drawn as lines between pairs of nodes, as illustrated in the following figure.

Formally an undirected Social Network/Graph can be defined as $G = (N, E)$, consisting of the set N of nodes and the set E of edges, which are unordered pairs of elements of N .

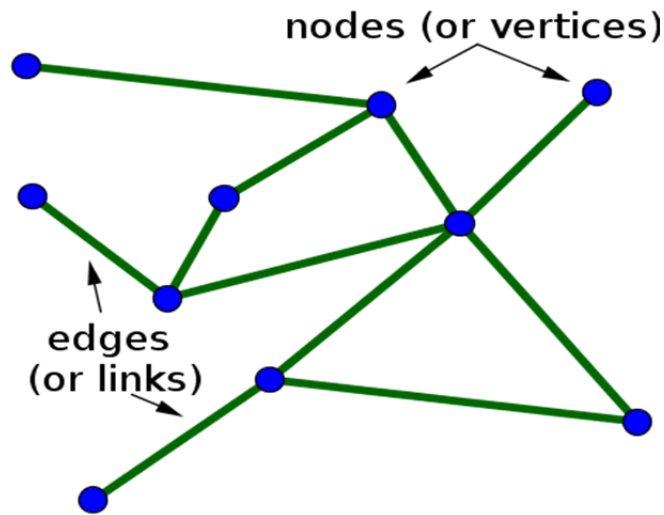


Figure 2.5 An undirected Social Network with 10 and 11 edges.

2.5 Social Network Analysis

Researches on social network analysis (SNA) stretches back more than half a century [10] where Jacob L. Moreno often is credited to be the researcher who was first systematically make use of social network analysis like techniques [11].

To define the social network analysis it focuses on the structure of relationships, ranging from casual acquaintance to close bonds. Social network analysis assumes that relationships are important. It maps and measures formal and informal relationships to understand what facilitates or impedes the knowledge flows that bind interacting units, viz., who knows whom, who work with whom and who shares what information and knowledge with whom by what communication media (e.g., data and information, voice, or video communications).[12] Social network analysis is a method with increasing application in the social sciences and has been applied in different area like link prediction from social networks, psychology, health, business organization, and electronic communications.

Social network analysis is considered in a vast amount of areas. To mention just a few, it can be used for understanding social interactions, to optimize flow of information between employees in a company, or to study and analyze criminal or terrorist organizations. Important problems within social network analysis are, among others:

1. To collect and extract useful data
2. To visualize the network in a way that support analysts with interpretation of the social structures.
3. To identify important structural patterns of the network (such as the identification of actors in the network that are extra important or powerful).
4. To predicate new links by calculating closeness, centralityetc.

Analyzing social networks enables us to detect several inter and intra connections between nodes in and outside their networks. The analysis of most social networks is an explanation of ties role of each node in the network. [13] The values explaining the role of nodes can be measured in the form of centrality values. There are 4 categories of centrality values namely degree centrality, closeness centrality, betweenness centrality, eigenvector centrality.

2.5.1 Basic Metrics in Social Network Analysis

2.5.1.1 Degree Centrality

Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has). In the directed network (meaning that ties have direction), we usually define two separate measures of degree centrality, namely in-degree and out-degree. In-degree is a count of the number of ties directed to the node, and out-degree is the number of ties that the node directs to others.

Formally for a graph $G \in (V, E)$ with n vertices, the degree centrality $C_D(v_i)$ for vertex v_i is as follow:

$$C_D(v_i) = \frac{d_i}{n-1} \quad (2.1)$$

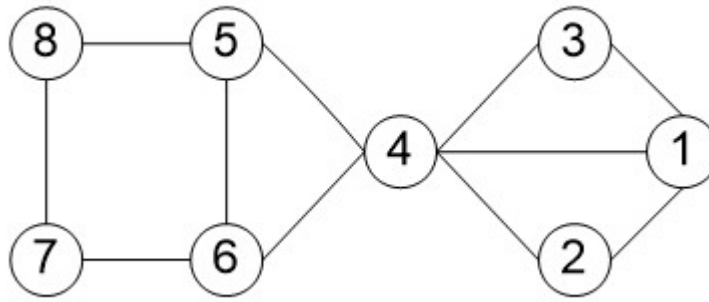


Figure 2.6 An example network for degree centrality calculation

$$C_D(4) = 5/7 = 0.71$$

$$C_D(7) = 2/7 = 0.28$$

In the figure 2.6, node 4 has the highest degree centrality while 7 and 8 have the lowest.

2.5.1.2 Closeness Centrality

Closeness centrality is the total number of spaces which an object has from other objects in the network, it shows the centralization of the under analysis object. Intuitively we say two sets are close if they are arbitrarily near to each other.

In graph theory closeness is a centrality measure of a vertex within a graph. Vertices that are 'shallow' to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. In the network theory, closeness is a sophisticated measure of centrality. It is defined as the mean geodesic distance (i.e., the shortest path) between a vertex v and all other vertices reachable from it.

Closeness centrality is the sum of all distances from an interesting node to other nodes in the network. It explains whether such a node is the center of the network or not. Formally for a graph $G \in (V, E)$ with n vertices, the closeness centrality C_C (v_i) for vertex v_i is as follow:

$$C_C(v_i) = \frac{n-1}{\sum_{j \neq i}^n g(v_i, v_j)} \quad (2.2)$$

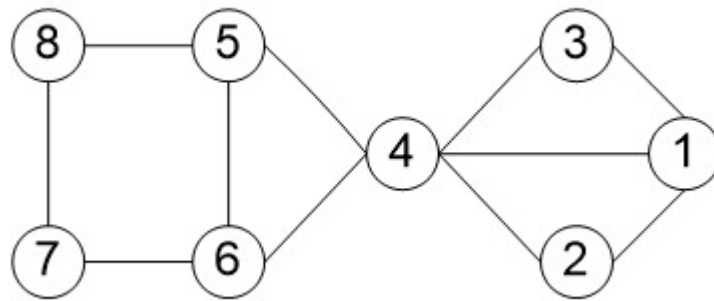


Figure 2.7 An example network for closeness centrality calculation

$$C_C(4) = 8 - 1/1 + 1 + 1 + 1 + 1 + 2 + 2 = 7/9 = 0.77 \quad C_C(5) = 8 - 1/1 + 1 + 1 + 2 + 2 + 2 + 3 = 7/12 = 0.58$$

In figure 2.7, node 4 is more central than 5.

2.5.1.3 Betweenness Centrality

Betweenness is a centrality metrics of a node within a graph (there is also edge betweenness, which is not discussed here). Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not. For a graph $G \in (V, E)$ with n vertices, the betweenness $C_B(v_i)$ for vertex v_i is computed as follows:

1. For each pair of vertices (s, t) , compute all shortest paths between them.
2. For each pair of vertices (s, t) , determine the fraction of shortest paths that pass through the vertex in question (here, vertex v_i).
3. Sum this fraction over all pairs of vertices (s, t) . Or, more succinctly:

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s < t} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \quad (2.3)$$

In the equation (2.3) $\sigma_{st}(v_i)$ is the number of shortest paths from s to t , and σ_{st} is the number of shortest paths from s to t that pass through a vertex v_i .

2.5.1.4 Eigenvector Centrality

Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Google's PageRank is a variant of the Eigenvector centrality measure.

2.5.1.5 Other Metrics

❖ **Bridge:**

Bridge is a link; if we delete it then the nodes which formed the mentioned link lie on different subgraphs.

❖ **Centralization:**

Centralization is the difference between the numbers of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the numbers of links each node possesses.

❖ **Clustering coefficient:**

A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.

❖ **Cohesion:**

Cohesion is the degree to which actors are connected directly to each other by cohesive bonds. Groups are identified as 'cliques' if every individual is directly tied to every other individual, 'social circles' if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.[20]

❖ **Degree:**

Degree is the sum of the links that an object/node has in a network. Definitely it is the sum of in-degree and out-degree of a node.

❖ **Density:**

The degree a respondent's ties know one another/ proportion of ties among an individual's nominees. Network or global-level density is the proportion of ties in a network relative to the total number possible (sparse versus dense networks).

❖ **Local bridge:**

Local bridge is an edge is a local bridge if its endpoints share no common neighbors. Unlike a bridge, a local bridge is contained in a cycle.

❖ **Path length:**

Path length is the distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.

❖ **Prestige:**

In a directed graph prestige is the term used to describe a node's centrality. "Degree Prestige", "Proximity Prestige", and "Status Prestige" are all measures of Prestige.

3. LINK PREDICTION

Link prediction focuses on finding hidden links or predicting connections that tend to become visible in the future time by considering the previous states of network [29] or this problem includes the new links prediction or finding invisible connections in a network. Link prediction is a significant task and it has a wide area of application, for instance citation networks, bioinformatics, e-commerce, recommending systems, co-authorship networks, etc. [48]. Link prediction is a significant task and it has a wide area of application, for instance citation networks, bioinformatics, e-commerce, recommending systems, co-authorship networks, bibliographic domain, criminal investigations and molecule biology [2], [4]. A classic definition of the link prediction problem is expressed by: “Given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t' ” [12]. Several approaches proposed to deal with this problem. Some well-known ones listed as: 1) Node-wise based approach 2) Topological/structural patterns based approach.

Extracting the values of measures that show the similarity between pairs of nodes is the basic point for node-wise based approach. Nodes act as a vector of features, in order to find their closeness, similarity metrics applied to each pairs of nodes. After that the scores can be used by unsupervised method [12], [17], [14]. In the unsupervised strategy, a proximity metric is select and apply to pairs of nodes in the network aiming to rank them. The top ranked pair of nodes are predict to make a tie in future.

The topological patterns based approach [13], [14], [15], [16] extract the scores from non-connected nodes of the network by using topological metrics. By tracking structural patterns of the network, these metrics provide a similarity degree between two nodes [16]. Then the scores used as a basis to build models for performing the prediction. Topological strategies are the most well-known among others. It is also easy to implement and presents good performance [13], [14], [15], [16].

In the most of previous studies the classic definition of link prediction has followed, the link prediction task performed by analyzing the complete structure of network at the current time and creation time of existing links or temporal information didn’t considered. However, temporal information (e.g., the interaction moments of nodes in the past or the first observation time of a

connection) is a meaningful perspective that should be focused during the link prediction task [18]. For instance, it would be helpful, if the formation time of links between neighbors not only how many times they interacted taken into account during the neighborhood based proximity score calculation. Common neighbor's recent activities can be more valuable than the old activities. In order to predict new connections they use the static structure of the network and changes in the network over the time didn't considered in these approaches, as a result they cannot model its evolution as such. Also, static approaches are suitable for investigating the occurrence of a certain link in a network but they are not so much useful, as an example, if the prediction of repeated link occurrences be a point of interest [19].

There are so many methods that deal with link prediction problem; the method which uses proximity metrics between two nodes is the most well-known among them [48].

Commonly, there are two types of approaches which deal with the prediction after the similarity scores are calculated:

3.1 Unsupervised Approach

The unsupervised strategy rank the non-connected pairs of nodes as a list, then the pairs of nodes which have the highest scores ordering in top of the list. Then the top listed pairs of nodes are predicted to be connected. This method is so easy to implement and does not require a labeled training set to perform the prediction. However, it shows some limitations such as the need to define the threshold L and the difficulty in combining information provided by more than one metric.

3.2 Supervised Approach

The link prediction problem is treated as a classification problem by supervised approach, in which pairs of nodes that are practically connected are assigned to class positive like (class 1), but the non-linked pairs of nodes are assigned to class negative like (class 0). Unlike the unsupervised method, the supervised approach needs a labeled training set to train the classifier that is going to be used. Finally, the prediction task can be performed by deploying a set of pairs of nodes in the trained classifier.

3.3 Related Works

In the literature of link prediction the most developed algorithms are based on topological information of network. According to the structural properties of the network, link prediction approaches can be categorized into three groups: 1) Common neighbor based 2) Global information based 3) Paths based.

Here we just give some information about common neighbor based methods. Common neighbor based methods take only information of first order neighbors into account namely, Common Neighbor Index [9], [6] Jaccard Index [47], [10] Preferential Attachment Index [28], [14] Adamic Adar Index [7] and Resource Allocation Index [48].

Although, most of link prediction studies didn't take weights of links into account, here we remark some previous studies in which the use of temporal information, weighted networks and citation networks in link prediction are the matters of importance. For instance, a weighted version of some unweighted proximity measures has proposed by Hially Rodrigues de Sa and Ricardo B. C. Prudencio in [14] and they record a good prediction performance with weighted proximity measures. Zhijie et al. proposed a method based on Markov Chain system in weighted networks with the use of Resource Allocation into it. Difference of nodes with various degrees and weights along various orders neighbors has considered by their method [32].

A weighted PageRank algorithm based on weighted directed author citation has proposed by Radicchi et al. in [33], aiming to rank the scientists by taking their scientific publication credits into account. In [18] a weighted graph modeled as a network by authors. Link's weights was acting as the age of the most recent activities among nodes and after that the extended version of Adamic Adar for weighted networks has deployed to accomplish the link prediction task. Mining of network data completed with temporal information for discovering the association rules which explain the network in a best way has proposed by authors in [21]. Juszczyszyn et al. acknowledged an approach in which the probabilities of transmissions between triads of nodes derived by using the history of the network (records of network during the past frames of time) [22]. In this approach compromising results can be achieved also a very expensive task which is frequent subgraphs mining has shown in this work [23]. In [19] a time series built for each pair of nodes by authors, in which the frequency of happening of links among the nodes during a particular cycle of time is the observation of each series. Potgieter in

[24] and Soares & Prudêncio in [25] adopted a similar idea, but according to their studies for prediction the proximity scores time series models were used. In this strategy, they used a chosen proximity metric and for each pair of nodes a time series was built by calculating the score in an array of time periods or various states of the network used over the time. As a result for each pair a final proximity score obtained by forecasting its corresponding time series.

3.4 Proximity Measures in Link Prediction

In the link prediction literature the evaluated and proposed proximity metrics can be grouped into semantic or topological measures [48]. In the node-wise metrics or semantic metrics for measuring the proximity nodes' information is taken in account. As an example, for predicting the oncoming interchanges between authors in a co-authorship net, the already extracted similarities among keywords which taken out from published articles can be used [48]. Unlike semantic metrics, in topological approach for computing the proximity values the structure of network is deploying. In link prediction literature topological metrics are more generally proposed than other approaches.

3.4.1 Number of Common Neighbors (CN)

The Number of common neighbors is one of the most widespread measures adapted to the link prediction problem which is so simple. This method senses that the future connection of two nodes is related to a high number of common neighbors between those nodes [14].

$$CN(i, j) = |\Gamma(i) \cap \Gamma(j)| \quad (3.1)$$

The number of common neighbors metric can be integrated for weighted networks as below:

$$CN(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} w(i, k) + w(j, k) \quad (3.2)$$

N. Contractor, M. Aurangzeb Ahmad ... considered the CN as Parametric Weighted Common Neighbors in [15] and they applied an (α) power to the neighbor nodes with this logic that “ $\alpha = 0$ brings the common neighbors metric and when $\alpha = 1$ it takes the reward of weights of the neighbors.” And they extended CN measure for weighted networks as:

$$S_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} w(i, k)^\alpha + w(k, j)^\alpha \quad (3.3)$$

T. Murata and S.Moriyasu have proposed a little bit different formula for CN in [16] and they are dividing the weights of common neighbors of a node in 2 as follow:

$$Score(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{2} \quad (3.4)$$

3.4.2 Jacquard's coefficient

The jacquard's coefficient measure is greatly discovered in data mining problems [17] for comparing the sets. This method presumes a high value for those pairs of nodes that share a high ration of common neighbors related to their total number of neighbors.

$$JC(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (3.5)$$

The jacquard's coefficient metric can be integrated for weighted networks as below:

$$JC(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{\sum_{a \in \Gamma(i)} w(a, i) + \sum_{b \in \Gamma(j)} w(b, j)} \quad (3.6)$$

3.4.3 Preferential Attachment (PA)

The preferential attachment metric assumes that the creation possibility of a connection between a pairs of nodes in future is rational to their shared degrees. (i.e, pairs of nodes which share a high number of relations at present will form new connections in future). Bonabeau and Barabasi [18], have suggested that the collaborators number between a pairs of nodes express the probability of a future link between them.

$$PA(i, j) = |\Gamma(i)| * |\Gamma(j)| \quad (3.7)$$

The preferential attachment metric can be integrated for weighted networks as below:

$$PA(i, j) = \sum_{a \in \Gamma(x)} w(a, i) * \sum_{b \in \Gamma(y)} w(b, j) \quad (3.8)$$

3.4.4 Adamic-Adar Coefficient (AA)

Adamic-Adar is related to Jaccard's coefficient and formulated by Adamic and Adar [19]. The common neighbors with fewer neighbors weighted/carried a higher importance in this metric. Moreover, AA evaluates the exclusiveness (or strength) of relationship between an evaluated pair of nodes and a common neighbor.

$$AA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log(\Gamma(k))} \quad (3.9)$$

The Adamic-Adar coefficient metric can be integrated for weighted networks as below:

$$AA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{\log(1 + \sum_{l \in \Gamma(k)} w(k, l))} \quad (3.10)$$

N. Contractor, M. Aurangzeb Ahmad ... proposed the below formula for AA as a Parametric Adar Adamic in [15].

$$S_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i,k)^\alpha + w(k,j)^\alpha}{\log(1 + s(k))} \quad (3.11)$$

T. Murata and S.Moriyasu have proposed the below formula for AA in [16]

$$Score(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{2} \times \frac{1}{\log(\sum_{k' \in \Gamma(k)} w(k', k))} \quad (3.12)$$

3.4.5 Path Distance (PD)

In un-weighted networks the path distance (PD) metric is given simply by calculation of minimal number of vertices which starts at i and follow till to reach to an end point j in the network. $PD(i, j) = l$ shows the pairs of nodes (i, j) that share a common neighbor. Creation chance of a new link between two nodes is related to lowness of the PD metric. In weighted networks this method take in account the minimal path between two nodes with consideration of $1/w(i, j)$ score which is the distance among neighbor nodes i and j .

3.4.6 Resource Allocation (RA)

Zhou et al. [21] proposed the RA matrix. Formally resource allocation is similar to adamic-adar coefficient (the exclusivity idea among a pair of nodes expressed by both methods), but their motivations are different from each other. Resource allocation's physical processes are the matter of importance in this method [20]. Resource Allocation is applicable to different kinds of networks for instance airport networks (passengers and aircraft's flow) and electric power station networks (power distribution).

$$RA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{|\Gamma(k)|} \quad (3.13)$$

The resource allocation metric can be integrated for weighted networks as below:

$$RA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{\sum_{l \in \Gamma(k)} w(k, l)} \quad (3.14)$$

N. Contractor, M. Aurangzeb Ahmad ... proposed the below formula for RA as a Parametric Resource Allocation in [15]

$$S_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k)^\alpha + w(k, j)^\alpha}{s(k)} \quad (3.15)$$

3.4.7 Local Path (LP)

Let $Paths_{x,y}^{(\ell)}$ presents a set of paths with a length of ℓ from i to j . In local path metric the exactly length of all paths like 2 or 3 among a pair of non-neighbors nodes are the matter of importance and this method counts all the paths [21]. Different from other methods which only analyse the connected neighbor's interactions, this metric has a wide range and takes in account the node's neighborhood information. In LP the paths with length 2 are more related than the paths with length 3, therefore a neighborhood factor e is applied to this metric.

$$LP(i, j) = |paths_{i,j}^{(2)}| + e * |paths_{i,j}^{(3)}| \quad (3.19)$$

Let i and j present the nodes under evaluation, this metric can be integrated for weighted networks as: for paths with length 2 $w(i, k) + w(j, k)$ is used, in which k is a common neighbor of i and j ; for paths with length 3 $w(i, m) + w(m, n) + w(n, j)$, in which m is a neighbor node to i but not to j , n is a neighbor node to j but not to i , finally m and n are connected directly. So local path metric for weighted networks can be given as below:

$$\begin{aligned}
 LP(i, j) = & \sum_{k \in \Gamma(i) \cap \Gamma(j)} w(i, k) + w(j, k) + e \\
 & * \sum_{\{i, m, n, j\} \in paths_{i, j}^{(3)}} w(i, m) + w(m, n) + w(n, j)
 \end{aligned} \tag{3.20}$$

4. LINK PREDICTION IN DIRECTED CITATION NETWORK

4.1 Directed Networks

When we are considering directed networks actually the directed graphs is going to be considered because each is representing a network. A directed Network (or digraph) (V, E) consists of a nonempty set of vertices V and a set of directed edges (or arcs) E . Each directed edge is associated with an ordered pair of vertices. The directed edge associated with the ordered pair (u, v) is said to start at u and end at v .

When depicting a directed network with a line drawing, using an arrow pointing from u to v indicate the direction of an edge that starts at u and ends at v . A directed network may contain loops and it may contain multiple directed edges that start and end at the same vertices. A directed network may also contain directed edges that connect vertices u and v in both directions; that is, when a digraph contains an edge from u to v , it may also contain one or more edges from v to u [48].

Directed graphs that may have multiple directed edges from a vertex to a second (possibly the same) vertex are used to model such networks. We called such graphs directed multi graphs. When there are m directed edges, each associated to an ordered pair of vertices (u, v) , we say that (u, v) is an edge of multiplicity m .

In Twitter, a connection between two users can establishes if a person is interested in news updates of another person. The link is thus directed because there is no mutual agreement needed to establish a link. The resulting network is directed and is also called follower network. User can follow an arbitrary number of other users to receive news or activity updates. The same in a citation network when an author is citing by someone the cited person or node receive a link or connection from opposite side and reverse; therefore the invented network is a directed network.

4.2 Background

The advance of internet gave this possibility to people and organizations to interact and collaborate more and more which provide the basis for the growth of social networks in the internet. Social networks are the outcome of nodes and edges that can be shown as a graph, in which the nodes act as organizations or people and edges act as different types of social relationships (such as citation, in which two authors are connected if they have cite each other) between them. In Social networks, ties/links tend to become visible, become stronger, become weaker and become invisible along the time that makes the dynamic of social networks.

Social Network Analysis (SNA) has a wide area of researches which tries to overpass these kinds of problems [1]. Several tasks can be related to SNA. In this paper our aim is to consider the dynamic of links in Author-Author social network, thus we want to predict those citation links that will form or become stronger along the different frames of time based on previous state of the network. SNA deal with such well-known problem called link prediction.

Lots of studies have been done to treat the link prediction problem [2], [3], [4]. The application of proximity metrics on the non-linked pairs of nodes at the current time in the network has been considered by most of studies in the past to predict new links at future time. These kinds of metrics give scores to any pair of nodes, and then the given scores use for performing the prediction task by an unsupervised or a supervised method. In unsupervised method, the non-connected pairs of nodes are rank by a chosen metric and then the top ranked pairs specify as a predicted links. In supervised method the link prediction problem consider as a classification task. In this method the connected pairs of nodes assign to positive class while non-connected ones assign to negative class. In this method the similarity scores which chosen from a set of topological metrics, accept as features and then use by a classifier for preforming the prediction task. In the previous studies the proximity score calculation usually done without taking the evolution of the network into account. This can be point as a limitation in the previous works. The proximity metrics computed using all network data up to the current time (i.e., present state of the network) without considering when links were created. Therefore, an effective source of data for link prediction is not taken into account.

In order to check the performance of our proposed approach, we performed experiments on Author-Author Citation networks that extracted from different parts of DBLP¹ (Digital Bibliography & Library Project). DBLP formed form bibliographic datasets which provides a big amount of data about various research publications in the Computer Science area. In order to experiment and compare our studies results with the already gained ones by traditional approaches, we used the common neighbors, Jacquard’s coefficient, preferential attachment and Adamic/Adar [6], [8], [7], [5] traditional similarity measures among others traditional similarity metrics in an unsupervised method [9], [14], [11].

4.3 Link Prediction

Link prediction focuses on finding hidden links or predicting connections that tend to become visible in the future time by considering the prior statuses of network. Link prediction is a significant task and it has a wide area of application, for instance citation networks, bioinformatics, e-commerce, recommending systems, co-authorship networks, bibliographic domain, criminal investigations and molecule biology [2], [4]. A classic definition of the link prediction problem is expressed by: “Given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t' ” [12]. Several approaches proposed to deal with this problem. Some well-known ones listed as: 1) Node-wise based approach 2) Topological/structural patterns based approach.

Extracting the values of metrics which show the similarity among two vertices is the basic point for node-wise based approach. Nodes act as a vector of features, in order to find their closeness, similarity metrics applied to each pairs of nodes. After that the scores can be used by unsupervised method [12], [17], [14]. In the unsupervised strategy, a proximity metric is select and apply to pairs of nodes in the network aiming to rank them. The top ranked pairs of nodes predict to make a tie in future.

Topological patterns based approach [13], [14], [15], [16] extract the scores from non-connected nodes of the network by using topological metrics. By tracking structural patterns of the network, these metrics provide a similarity degree between two nodes [16]. Then the scores used as a basis to build models for performing the prediction. Topological strategies are the most

well-known among others. It is also easy to implement and presents good performance [13], [14], [15], [16].

In the most of previous studies the classic definition of link prediction has followed, the link prediction task performed by analyzing the complete structure of network at the current time and creation time of existing links or temporal information didn't considered. Nevertheless, temporal information (like the interaction moments of vertices in the past or the first observation time of a tie/link) is a meaningful perspective which should be focused in the link prediction task [18]. For instance, it would be helpful, if the formation time of links between neighbors not only how many times they interacted taken into account during the neighborhood based proximity score calculation. Common neighbor's recent activities can be more valuable than the old activities. In order to predict new connections they use the static structure of the network and changes in the network over the time didn't considered in these approaches, as a result they cannot model its evolution as such. Also, static approaches are suitable for investigating the occurrence of a certain link in a network but they are not so much useful, as an example, if the prediction of repeated link occurrences be a point of interest [19].

4.4 Used Traditional Proximity Measures

In this section, we explained the traditional measures applied as predictor and comparator attributes in our unsupervised link prediction approach.

4.4.1 Number of Common Neighbors (CN)

The most widespread and simplest metric acknowledged in link prediction problem is the common neighbor (CN) metric. Also, a high number of common neighbors make it possible that a connection between two node x and y will create in future [6]. The CN metric for un-weighted networks is given as follow:

$$CN(i, j) = |\Gamma(i) \cap \Gamma(j)| \quad (4.1)$$

4.4.2 Jacquard's coefficient

For comparing sets in data mining the JC metric is well explored [47]. The JC presume the likelihood of creation of a new connection among a pair of nodes that share a high ratio of common neighbors comparative to the total number of their adjacent nodes. Jacquard's coefficient for un-weighted networks is given as:

$$JC(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (4.2)$$

4.4.3 Preferential Attachment (PA)

Preferential attachment metric supposes that the occurrence probability of a future connection among a pair of nodes is rational to their shared degrees. (i.e., nodes that share a high amount of relations at the present will form new connection future). Bonabeau and Barabasi [28], have suggested that production of collaborators number among a pair of nodes can express the probability of a future link between them. The PA metric for un-weighted networks is given as below:

$$PA(i, j) = |\Gamma(i)| * |\Gamma(j)| \quad (4.3)$$

4.4.4 Adamic-Adar Coefficient (AA)

This measure is formulated by Adamic and Adar [7] which is relevant to Jaccard's coefficient. The common neighbors with fewer neighbors weighted/carried a higher importance in this metric. Moreover, AA evaluates the exclusiveness (or strength) of relationship between an evaluated pair of nodes and a common neighbor. The AA measure is given for un-weighted networks as below:

$$AA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log(\Gamma(k))} \quad (4.4)$$

4.5 Citation Network and Triad Patterns

Dyad is the basic unit of analysis in social network theory. In undirected networks, a dyad is a pair of nodes who may share a social relation with one another. In directed networks, a dyad is a pair of nodes who may share a social relation through mutual links, a nonreciprocal relation, or no relation. Nonreciprocal means that one node is interested in the other node but the other node is not interested. A set of three parties, which includes three dyads, is called triad. A triad is ‘closed’ if all nodes are linked with each other in some manner. A closed triad is also called triangle. Figure 4.1 shows triad patterns of the actors x , z , and y . Edges are directed because our aim is to model patterns in directed social networks (e.g., citation networks). All patterns are open triads with z being the common neighbor of x and y [29].

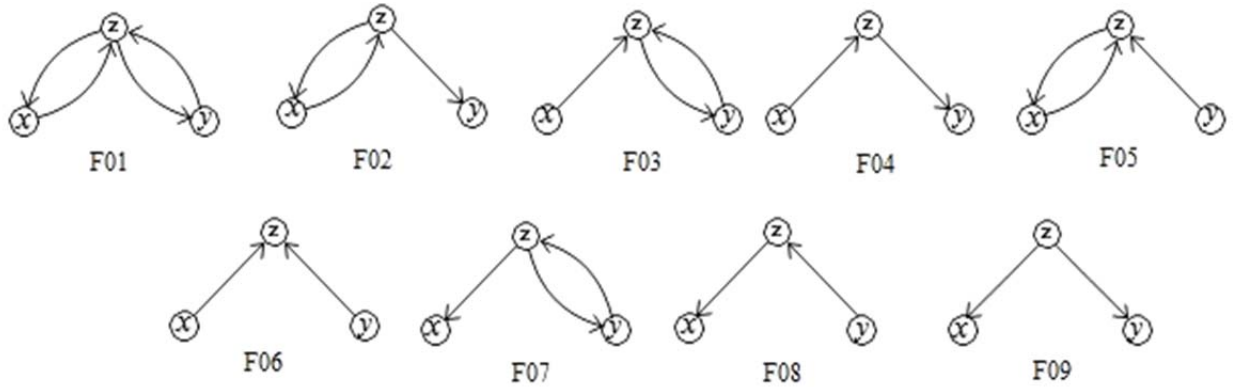


Figure 4.1 Triad graph patterns

Figure (4.1) shows all possible connectivity configurations between x , z , and y with the condition that x and y are not directly connected. Open triads are labeled as $F0X$ where X is a changing index between [1, 9]. The pattern $F01$ shows the case where x and z as well as y and z are mutually connected. According to the theory of triadic closure, the chances are high that x will also connect to y (i.e., x and z as well as y and z mutually cited each other, x and y will possibly cite each other also). In $F02$, only x and z are mutually cited by each other. The node y is cited by z but the relationship is not reciprocated. $F03$, $F05$, and $F07$ indicate

complementary cases where a mutual citation among one dyad exists. The other cases labeled as $F04$, $F06$, $F08$, and $F09$ shows patterns without any mutual citation among the dyads. The goal which follows by link prediction is to determine which one of the triads are or will be closed (i.e., becoming a triangle). A triad can be closed as follow, if x cites y , y cites x , or if x and y mutually cite each other. The figures below shows closed triads based on $F01$ and $F09$ (the first and the last pattern of Figure (4.1) are shown for brevity).

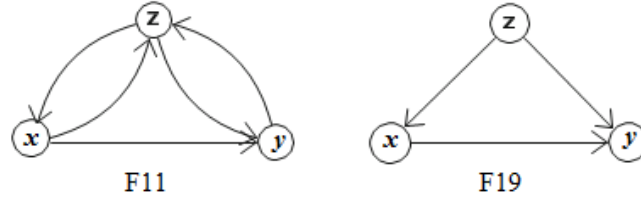


Figure 4.2 Closed F1X triads

Figure (4.2) shows the patterns where the triads are closed from x to y . The triads which are closed via x to y are labeled as $F1X$ with $X = [1, 9]$. In the same manner, the triads that are closed via y to x are labeled as $F2X$ with $X = [1, 9]$ which is shown in figure (4.3).

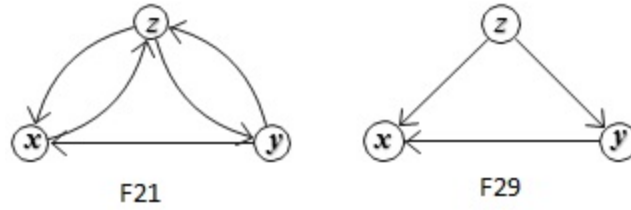


Figure 4.3 Closed F2X triads

Finally, the triads which are closed by mutual connections of x and y (figure 4.4), labeled as $F3X$ with $X = [1, 9]$.

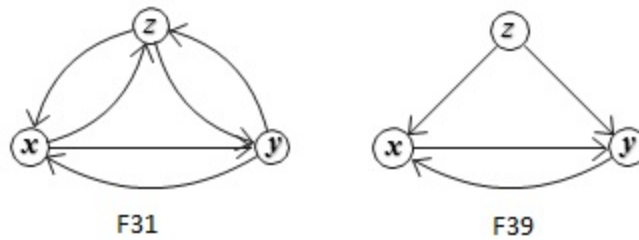


Figure 4.4 Closed F3X triads

To summarize the discussions about triad patterns, triads that are related to link prediction may have 36 different configurations with respect to how nodes are connected to each other through directed links. Open triads have 2 connected dyads and have 2 to 4 links. Closed triads have 3 connected dyads and have 3 to 6 links.

4.6 The Proposed Time-Frame Based Link Prediction in Directed Networks

4.6.1 Temporal Structure

Let's observe the network G at time t that should be split into numerous time-segment snapshots, which present the network at different times in the past. Then a prediction window defined, that shows how further we want to make the prediction in the future. Afterword, small sets which we named them frames grouped in the sequential snapshot which includes snapshots according to the length of prediction window.

Let G_t represent the network at time t . The frame created from the union of the graphs from time 1 to T shows by $[G_1, G_2, \dots, G_T]$. The number of periods (frames) in the series given by n . And w presents the prediction window.

In this paper we focused on a network which indicate users information up to the year $T = 2012$ with a 4 year length of prediction window (i.e., new links prediction from 2009 up to 2012 is the task). We extracted $k = 3$ frames from the network structure and N obtained as $(N = F1, F2, F3)$.

$$N = \{[1997-2000], [2001-2004], [2005-2008], [2009, 2012]\}$$

4.6.2 Temporal event

A temporal event occurs when the states of two vertices changes from a status like linked or non-linked to other state along the different frame of time within evolves of network from a frame to another frame. Three types (Protective, Inventive or Regressive) of events has categorized below.

4.6.2.1 Protective

A protective event occurs when a dyad's relation is not dropped with the evolvement of the network that is when a dyad share a connection in a frame and the link is remain in the next frame. With respect to the above mentioned graph patterns protective event can be redefined as, a protective event occur when a triad safe its states without any changes along the evolvement of the network from a frame to the other one. In the formula below $P(i, j, k)$ is the reward of a protective event for (i, j) nodes in the frame F_k which is the matter of importance within the evolution of network between $(k-1)$ th frame to the k th ones:

$$p(i, j, k) = \begin{cases} p, & \text{if } n_{k-1}(i, j) = n_k(i, j) \\ 0, & \text{otherwise} \end{cases} \quad (4.5)$$

In the equation (4.5), $n_{k-1}(i, j)$ is the number of edges between i and j nodes in the $(k-1)$ -th frame. If there is a mutual relation between the related nodes, this number equals to 2. The constant p shows the reward for protective event, as the link between two nodes is stable the event's value should be a positive. E_{k-1} and E_k present sets of edges detected in frames F_{k-1} and F_k one by one.

4.6.2.2 Inventive

The formation of new connection among pairs of nodes in various frames of time is showed by inventive event. This happen when there is no connection between two nodes in a frame and a connection appears in the following frame. In the formula below $I(i, j, k)$ is the reward of an inventive event for (i, j) nodes in the frame F_k which is the matter of importance within the evolution of network between $(k-1)$ th frame to the k th ones:

$$I(i, j, k) = \begin{cases} I, & \text{if } n_k(i, j) > n_{k-1}(i, j) \\ 0, & \text{otherwise} \end{cases} \quad (4.6)$$

In the above equation the constant I indicates the reward for inventive events. Since the tie between two nodes becomes strong its value should be positive.

4.6.2.3 Regressive

Regressive events are the opposite form of inventive events. The elimination of an existent connection among pairs of nodes in two different frames is representing by regressive event. In the formula below $r(i, j, k)$ is the reward of a regressive event for (i, j) nodes in the frame F_k which is the matter of importance within the evolution of network between $(k-1)$ th frame to the k th ones:

$$r(i, j, k) = \begin{cases} r, & \text{if } n_{k-1}(i, j) > n_k(i, j) \\ 0, & \text{otherwise} \end{cases} \quad (4.7)$$

In the above equation the constant r indicates the reward for regressive events. Since the tie between two nodes tends to decrease then its value should be negative.

Figure (4.5) shows an example of the above mentioned three events (Protective, Inventive & Regressive) in which, we can see that a protective event occurred between node z & y after the evolvement of the network from frame $k-1$ to frame k . A link reduction can be seen between x and z which is the representation of the regressive event on other hand there is no connection between x and y in frame $k-1$ while a new link observed in frame k between them which present the inventive event.

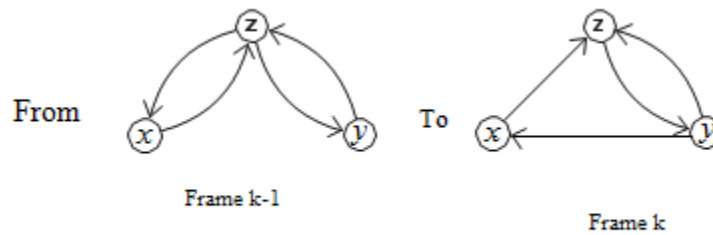


Figure 4.5 Protective, inventive and regressive events

4.7 Time Frame Based Score

There are different kinds of strategies for link prediction that calculate scores to pair of nodes by applying proximity metrics, to find how similar are the vertices and is there any connection tend to appear or form between them in the future.

In the proposed method we use the rewards of secondary events, that shows the consideration of temporal events in nodes' neighborhood, not the one directly related to the under analysis pairs of nodes.

$$Score(i, j) = \sum_{k=2}^n P(i, j, k) + \alpha \times s(i, j, k) \quad (4.8)$$

$$P(i, j, k) = \begin{cases} p(i, j, k) & \text{if protective event observed} \\ I(i, j, k) & \text{if inventive event observed} \\ r(i, j, k) & \text{if regressive event observed} \end{cases} \quad (4.9)$$

$$S(i, j, k) = \sum_{l \in \Gamma(i) \cap \Gamma(j)} P(i, l, k) + P(j, l, k) \quad (4.10)$$

In equation (4.8) α is parameter that shows the strong affection of secondary events over the link between nodes i and j . In Equation (4.9) the reward of (Protective, Inventive or Regressive) event for (i, j) pairs, which considered in the transition from frame $k-1$ to frame k has calculated by $P(i, j, k)$. In Equation (4.10) $S(i, j, k)$ shows the collected rewards of secondary events appointed to the pairs (i, j) and the primary events considered in nodes common neighborhood, $\Gamma(i)$ is the set of neighbors of node i in the network.

In our experiment, the proposed metrics have compared with commonly used metrics in the literature for link prediction. For this purpose, Common Neighbors (CN) [6], Jaccard's Coefficient (JC) [47], Preferential Attachment (PA) [28] and Adamic-Adar Coefficient (AA) [7] metrics has been used.

The steps of the proposed method are as follows:

1. The non-connected pairs are selected from the frame before the prediction frame (validation frame). The non-connected pairs that didn't meet with any events during the network evolution were eliminated.
2. The proximity scores of five different predictors (CN, JC, PA, AA, and our method) are computed for the chosen pairs.
3. The ranked top n -pairs are taken into account as the future links that are more likely to appear.
4. If a non-connected pair in the validation set converts a connected pair in the prediction frame, the pair of nodes is assumed as positive and as negative otherwise.
5. The performance measures of all the predictors considered are compared.

4.8 Experimental Results

In order to evaluate the performance of the proposed method we conducted a set of experiments using DBLP-Citation-Network-V7 dataset. We downloaded the dataset from Arnetminer (<https://aminer.org>) which already extracted from DBLP and other sources by providers [34]. The dataset contains 2,244,021 papers and 4,354,534 citation relationships. We used the validation set to empirically evaluate and determine the most appropriate values of p , i , r and α parameters in the link prediction task. We showed with some tests that the best performance was achieved at $p = 0.40$, $i = 0.80$, $r = -0.37$ and $\alpha = 0.1$. These results indicate that protective (p) and regressive (r) values almost balance with each other. The best performance values were achieved when $p < i$.

The first experiment is dedicated to the evaluation of precision values for the proposed method and the traditional proximity scores on the network with $\{[1997-2000], [2001-2004], [2005-2008], [2009-2012]\}$ frames. In this structure, the frame size is 4 and the snapshot at $[2009-2012]$ time interval is used as the prediction frame. Precision values found by five methods are reported in Table 1. Precision can be evaluated at different points in a ranked list of extracted citations. Mathematically, precision at rank n ($P@n$) is defined as the proportion of relevant citations and extracted citations.

$$P@n = \frac{\text{number of relevant citations with rank } n}{n} \quad (4.11)$$

Table 1 Precision values (%) of five methods

Method	P@10	P@50	P@100	P@500
CN	0	4	21	38
JC	10	12	29	41
PA	10	21	28	53
AA	50	59	64	68
Our method	70	78	81	84

As can be easily seen from Table 1, our approach outperforms the other well-known predictors for every four P@n. Our method also performs best at Precision@500. CN delivers worst in terms of P@n.

5. LINK PREDICTION IN WEIGHTED CITATION NETWORKS

5.1 Weighted Networks

A weighted network is a network where the ties among nodes have weights assigned to them. A network is a system whose elements are somehow connected the elements of a system are represented as nodes (also known as actors or vertices) and the connections among interacting elements are known as ties, edges, arcs, or links. The nodes might be neurons, individuals, groups, organizations, airports, or even countries, whereas ties can take the form of friendship, communication, collaboration, alliance, flow, or trade, to name a few.

In a number of real-world networks, not all ties in a network have the same capacity. In fact, ties are often associated with weights that differentiate them in terms of their strength, intensity, or capacity [44] On the one hand, Mark Granovetter et al., [45] argued that the strength of social relationships in social network is a function of their duration, emotional intensity, intimacy, and exchange of services. On the other, for non-social networks, weights often refer to the function performed by ties

By recording the strength of ties, a weighted network can be created (also known as a valued network). Below is an example of such a network:

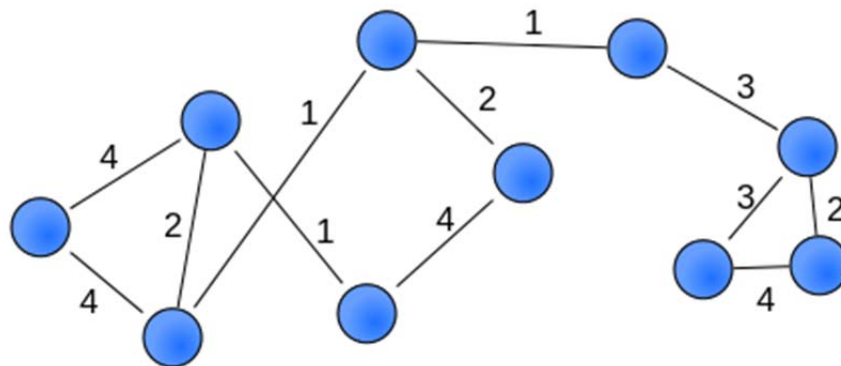


Figure 5.1 A simple weighted network

5.2 Measures for Weighted Networks

Although weighted networks are more difficult to analyze than if ties were simply present or absent, a better way to analyze weighted networks would be to redefine and generalize current methods to explicitly take weights into account. Currently, only a small set of measures have been generalized a number of network measures has been proposed for weighted networks:

- Node strength: The sum of weights attached to ties belonging to a node.
- Closeness: Redefined by using Dijkstra's distance algorithm.
- Betweenness: Redefined by using Dijkstra's distance algorithm.
- The clustering coefficient (global): Redefined by using a triplet value.
- The clustering coefficient (local): Redefined by using a triplet value or using an algebraic formula

A theoretical advantage of weighted networks is that they allow one to derive relationships among different network measures (also known as network concepts, statistics or indices. For example, Dong and Horvath et al, [25] show that simple relationships among network measures can be derived in clusters of nodes (modules) in weighted networks. For weighted correlation networks, one can use the angular interpretation of correlations to provide a geometric interpretation of network theoretic concepts and to derive unexpected relationships among them.

5.3 Link Prediction

Link prediction focuses on finding hidden connections or predicting links that tend to appear in the future time by considering the previous states of network. It is a very well-known task applicable to a broad of areas, such as citation networks, recommending systems, bibliographic domain, criminal investigations and molecule biology [2, 4]. Classical description of link prediction problem is stated by: “Given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t' ” [24]. Different approaches proposed to deal with this problem. Some well-known ones listed as: 1) Node-wise based approach 2) Topological/structural patterns based approach.

Extracting the values of metrics which show the similarity among two vertices is the basic point for node-wise based approach. Nodes act as a vector of features, in order to find their closeness similarity metrics applied to each pairs of nodes. After that the scores can be used by unsupervised method [11, 24-27]. In the unsupervised strategy, a proximity measure is selected and applied to pairs of nodes in the network aiming to rank them. The top ranked pairs of nodes predict to form a link in future.

Topological patterns based approach [9, 11, 28, 29] extract the scores from non-connected nodes of the network by using topological metrics. By tracking the network's structural patterns, these metrics provide a similarity degree among a pair of nodes [9]. Then the scores used as a basis to build models for performing the prediction. The topological based approach is the most widespread one. It also easy to implement and presents good performance.

In the most of previous studies the classic definition of link prediction has followed, the link prediction task performed by analyzing the complete structure of network at the current time and creation time of existing links or temporal information didn't considered. Nevertheless, temporal information (like interaction moments of vertices in the past or the first observation time of a tie/link) is a meaningful perspective which should be focused in the link prediction task [30-31]. For instance, it would be helpful, if the formation time of links between neighbors not only how many times they interacted taken into account during the neighborhood based proximity score calculation. Common neighbor's recent activities can be more valuable than old activities. In order to predict new connections they use the static structure of the network and changes in the network over the time don't considered in these approaches, as a result they cannot model its evolution as such. Also, static approaches are suitable for investigating the occurrence of a certain link in a network but they are not so much useful, as an example, if the prediction of repeated link occurrences be a point of interest.

5.4 Traditional proximity measures

In this section, we explained the traditional measures applied as predictor and comparator attributes in our unsupervised link prediction approach.

5.4.1 Number of Common Neighbors (CN)

The most widespread and simplest metric acknowledged in link prediction problem is the common neighbor (CN) metric. Also, a high number of common neighbors make it possible that a connection between two node x and y will create in future [6]. The CN metric for un-weighted networks is given as follow:

$$CN(i, j) = |\Gamma(i) \cap \Gamma(j)| \quad (5.1)$$

The weighted version of common neighbor (CN) metric can be extended as below for weighted networks.

$$CN(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x, z) + w(y, z) \quad (5.2)$$

5.4.2 Jacquard's coefficient

For comparing sets in data mining the JC metric is well explored [47]. The JC presume the likelihood of creation a new link between two nodes that share a high ratio of common neighbors comparative to the total number of their adjacent nodes. Jacquard's coefficient for un-weighted networks is given as:

$$JC(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (5.3)$$

The weighted version of Jacquard's coefficient (JC) metric can be extended as below for weighted networks.

$$JC(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{\sum_{a \in \Gamma(x)} w(a, x) + \sum_{b \in \Gamma(y)} w(b, y)} \quad (5.4)$$

5.4.3 Preferential Attachment (PA)

Preferential attachment metric supposes that the occurrence probability of a forthcoming connection among a pair of nodes is rational to their shared degrees. (i.e., nodes which have a high amount of relations at the present will form new connections in future). Newman [14] Bonabeau and Barabasi [28], have suggested that production of collaborators number among a pair of nodes can express the probability of a future link between them. Preferential attachment for un-weighted networks is given as below:

$$PA(i, j) = |\Gamma(i)| * |\Gamma(j)| \quad (5.5)$$

The weighted version of preferential attachment (PA) metric can be extended as below for weighted networks.

$$PA(i, j) = \sum_{m \in \Gamma(i)} w(m, i) * \sum_{n \in \Gamma(j)} w(n, j) \quad (5.6)$$

5.4.4 Adamic-Adar Coefficient (AA)

This measure which is relevant to Jaccard's coefficient have formulated by Adamic and Adar [7]. The common neighbors with fewer neighbors weighted/carried a higher importance in this metric. Moreover, AA evaluates the exclusiveness (or strength) of relationship between an evaluated pair of nodes and a common neighbor. The AA measure is given for un-weighted networks as below:

$$AA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log(\Gamma(k))} \quad (5.7)$$

The weighted version of Adamic and Adar (AA) metric can be extended as below for weighted networks.

$$AA(i, j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{w(i, k) + w(j, k)}{\log(1 + \sum_{l \in \Gamma(k)} w(k, l))} \quad (5.8)$$

5.5 Citation network and triad patterns

A directed weighted graph in which, each node indicates an author and each edge indicates a citation link from an author to another one, and weights show the citation numbers between two nodes. This can be shown by an arrow (link) going from the node showing A_i to the node showing A_j . In this case the nodes from a collection A (authors) form a directed weighted citation network. The figure below represents a citation network over the time.

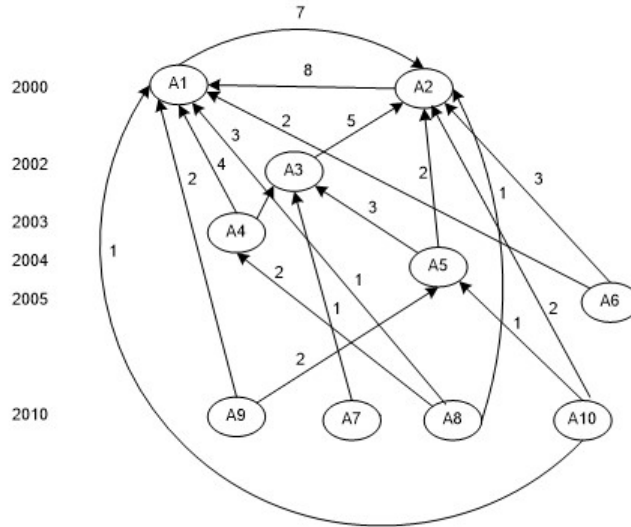


Figure 5.2 Citation network

Dyad is the basic unit of analysis in social network theory. In undirected networks, a dyad is a pair of nodes who may share a social relation with one another. In directed networks, a dyad is a pair of nodes who may share a social relation through mutual links, a nonreciprocal relation. Nonreciprocal means that one node is interested in the other node but the other node is not interested. A set of three parties, which includes three dyads, is called triad.

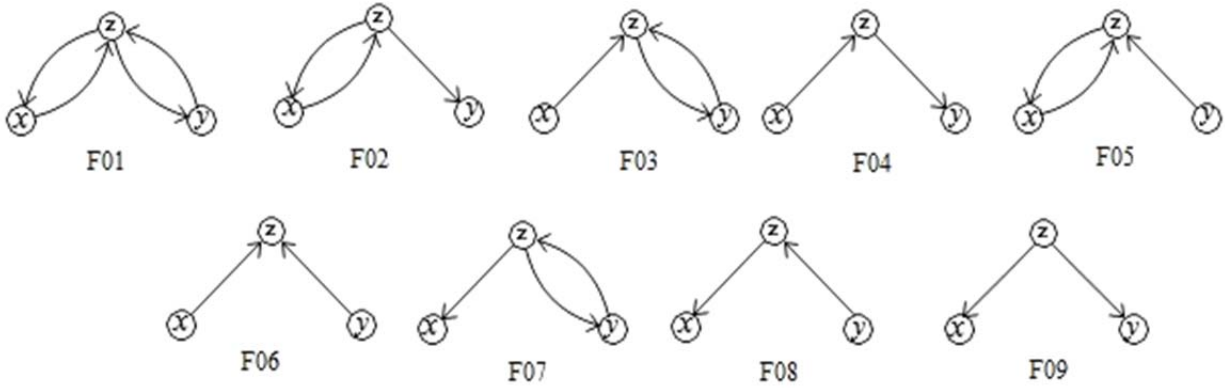


Figure 5.3 Triad graph patterns

A triad is ‘closed’ if all nodes are linked with each other in some manner. A closed triad is also called triangle. Figure (5.3) shows triad patterns of the actors x , z , and y .

Edges are directed because our aim is to model patterns in directed social networks (e.g., citation networks). All patterns are open triads with z being the common neighbor of x and y [29].

Figure (5.3) shows all possible connectivity configurations between x , z , and y with the condition that x and y are not directly connected. Open triads are labeled as $F0X$ where X is a changing index between [1, 9]. The pattern $F01$ shows the case where x and z as well as y and z are mutually connected. According to the theory of triadic closure, the chances are high that x will also connect to y (i.e., x and z as well as y and z mutually cited each other, x and y will possibly cite each other also). In $F02$, only x and z are mutually cited by each other. The node y is cited by z but the relationship is not reciprocated. $F03$, $F05$, and $F07$ indicate complementary cases where a mutual citation among one dyad exists. The other cases labeled as $F04$, $F06$, $F08$, and $F09$ shows patterns without any mutual citation among the dyads. The goal which follows by link prediction is to determine which one of the triads are or will be closed (i.e., becoming a triangle). A triad can be closed as follow, if x cites y , y cites x , or if x and y mutually cite each other. The figures below shows closed triads based on $F01$ and $F09$ (the first and the last pattern of Figure (5.3) are shown for brevity). Figure (5.4) shows the patterns where the triads are closed from x to y .

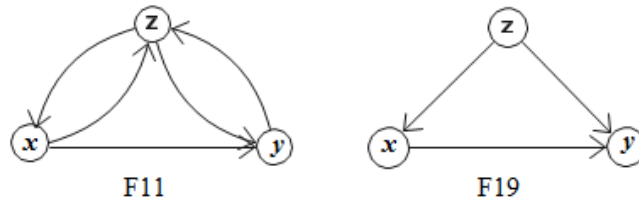


Figure 5.4 Closed F1X triads

The triads which are closed via x to y are labeled as $F1X$ with $X = [1, 9]$. In the same manner, the triads that are closed via y to x are labeled as $F2X$ with $X = [1, 9]$ which is shown in figure (5.5). Finally, the triads which are closed by mutual connections of x and y (Figure 5.6), labeled as $F3X$ with $X = [1, 9]$.

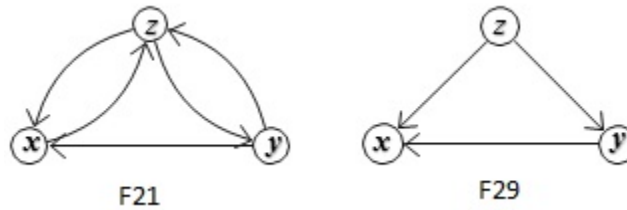


Figure 5.5 Closed F2X triads

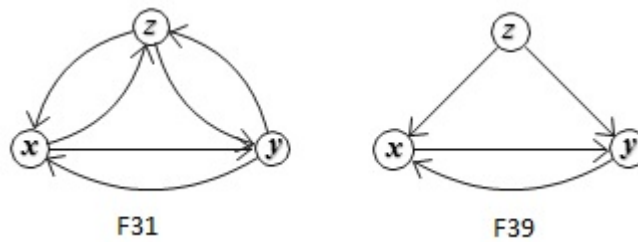


Figure 5.6 Closed F3X triads

To summarize the discussions about triad patterns, triads that are related to link prediction may have 36 different configurations with respect to how nodes are connected to each other through directed links. Open triads have 2 connected dyads and have 2 to 4 links. Closed triads have 3 connected dyads and have 3 to 6 links.

5.6 The Proposed Time-Frame Based Link Prediction in Directed Networks

5.6.1 Temporal Structure

Generally the prediction of new links with a particular type between a pair of nodes $x, y \in V$ in the future is the link prediction task in social networks. Temporal link prediction considers the dynamic evolvement of network, different from static link prediction. Prediction of links at time $T + 1$ is refers to temporal link prediction by using the links and nodes in time T . Let's observe the network G at time t that should be split into numerous time-segment snapshots, which present the network at different times in the past. Then a prediction window defined, that shows how further we want to make the prediction in the future. Afterword, small sets which we named them frames grouped in the sequential snapshot which includes snapshots according to the length of prediction window.

Let G_t represent the network at time t . The frame created from the union of the graphs from time 1 to T shows by $[G_1, G_2, \dots, G_T]$. The number of periods (frames) in the series given by n . And w presents the prediction window.

In this paper we focused on a network which indicate users information up to the year $T = 2012$ with a 4 year length of prediction window (i.e., new links prediction from 2009 up to 2012 is the task). We extracted $k = 3$ frames from the network structure and N obtained as $(N = F1, F2, F3)$.

$$N = \{[1997-2000], [2001-2004], [2005-2008], [2009, 2012]\}$$

5.6.2 Temporal event

A temporal event occurs when the states of two vertices changes from a status like linked or non-linked to other state along the different frame of time within evolves of network from a frame to another frame. Three types (Protective, Inventive or Regressive) of events has categorized below.

5.6.3 Weighted temporal event

A weighted temporal event occurs when the state of a link changes from a state (connected, non-connected, became stronger or became weaker) to another state or an e changes appear in link's weight along the different frame of time within evolves of network from a frame to another frame. Three types (W-Protective, W-Inventive and W-Regressive) of events has categorized below.

5.6.3.1 W-Protective

A weighted protective event occurs when total weight of relations between two nodes does not decrease less than an event changing rate (e) pre-determined or does not increase more than this rate through the evolvement of network from a time frame to another one.

In order to take a weighted protective event into account within the evolution from frame $(k-1)$ -th to the k -th, $w-p(u, v, k)$ reward for each pair of nodes (u, v) related to frame F_k is defined as follows:

$$w - p(i, j, k) = \begin{cases} p & \text{if } (w_{k-1}(i, j) + w_{k-1}(j, i)) - e \cdot (w_{k-1}(i, j) + w_{k-1}(j, i)) < \\ & (w_k(i, j) + w_k(j, i)) < \\ & (w_{k-1}(i, j) + w_{k-1}(j, i)) + e \cdot (w_{k-1}(i, j) + w_{k-1}(j, i)) \\ 0 & \text{otherwise} \end{cases} \quad (5.9)$$

In the equation (5.9), the constant p shows the reward for weighted protective event, $w_{k-1}(i, j)$ and $w_{k-1}(j, i)$ are the weights of edges between i and j , and j and i in the frame F_{k-1} , respectively. e ($0 < e \leq 1$) is the event changing rate pre-determined, the events between frames changes with respect to this rate.

5.6.3.2 W-Inventive

A weighted inventive event occurs when total weight of relations between two nodes increases more than the event changing rate (e) in the next frame. For each pair of nodes (i, j) an $wi(i, j, k)$ reward related to frame F_k has defined, in order to take a weighted inventive event into account within evolution from the $(k-1)th$ to the kth frame and its formula is as follow:

$$w - I(i, j, k) = \begin{cases} i & \text{if } (w_k(i, j) + w_k(j, i)) > \\ & (w_{k-1}(i, j) + w_{k-1}(j, i)) + e. (w_{k-1}(i, j) + w_{k-1}(j, i)) \\ 0 & \text{otherwise} \end{cases} \quad (5.10)$$

In the equation above the constant I indicates the reward for weighted inventive events. Since the tie between two nodes becomes stronger its value should be positive and $wEk-1$ and wEk are the sets of edges observed in frames $Fk-1$ and Fk respectively.

5.6.3.3 W-Regressive

A weighted regressive event occurs when total weight of relations between two nodes decreases more than the event changing rate (e) in the subsequent frame. For each pair of nodes (i, j) an $wr(i, j, k)$ reward related to frame F_k has defined, in order to take a weighted regressive event into account within the evolution from the $(k-1)th$ to the kth frame and its formula is as below:

$$w - r(i, j, k) = \begin{cases} r & \text{if } (w_k(i, j) + w_k(j, i)) < \\ & (w_{k-1}(i, j) + w_{k-1}(j, i)) - e. (w_{k-1}(i, j) + w_{k-1}(j, i)) \\ 0 & \text{otherwise} \end{cases} \quad (5.11)$$

In the equation (44) the constant wr indicates the reward for weighted regressive events. Since the tie between two nodes tends to decrease then its value should be negative.

Figure (5.7) shows an example of the above mentioned three events (Protective, Inventive & Regressive) in which, we can see that a protective event occurred between node z & y after the evolvement of the network from frame $k-1$ to frame k . A link reduction can be seen between x and z which is the representation of the regressive event on other hand there is no connection between x and y in frame $k-1$ while a new link observed in frame k between them which present the inventive event.

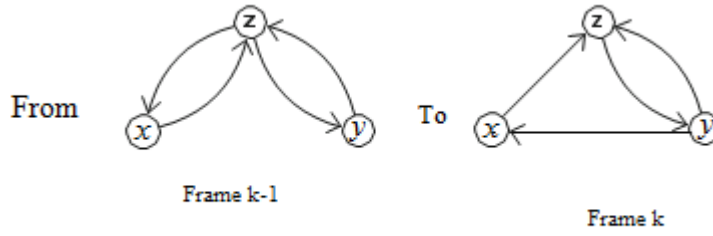


Figure 5.7 Protective, inventive and regressive events

5.7 Time Frame-based Score

There are different kinds of strategies for link prediction that calculate scores to pair of nodes by applying proximity metrics, to find how similar are the vertices and is there any connection tend to appear or form between them in the future.

In the proposed method we use the rewards of secondary events, that shows the consideration of temporal events in nodes' neighborhood, not the one directly related to the under analysis pairs of nodes.

$$Score(i, j) = \sum_{k=2}^n P(i, j, k) + \alpha \times s(i, j, k) \quad (5.12)$$

$$P(i, j, k) = \begin{cases} p(i, j, k) & \text{if protective event observed} \\ I(i, j, k) & \text{if inventive event observed} \\ r(i, j, k) & \text{if regressive event observed} \end{cases} \quad (5.13)$$

$$S(i, j, k) = \sum_{l \in \Gamma(i) \cap \Gamma(j)} P(i, l, k) + P(j, l, k) \quad (5.14)$$

In equation (5.12) α is parameter that shows the strong affection of secondary events over the link between nodes i and j . In equation (5.13) the reward of (Protective, Inventive or Regressive) event for (i, j) pairs, which considered in the transition from frame $k-1$ to frame k has calculated by $P(i, j, k)$. In equation (4.14) $S(i, j, k)$ shows the collected rewards of secondary events appointed to the pairs (i, j) and the primary events considered in nodes common neighborhood, $\Gamma(i)$ is the set of neighbors of node i in the network.

In our experiment, the proposed metrics have compared with commonly used metrics in the literature for link prediction. For this purpose, Common Neighbors (CN) [6], Jaccard's Coefficient (JC) [47], Preferential Attachment (PA) [28] and Adamic-Adar Coefficient (AA) [7] metrics has been used.

The steps of the proposed method are as follows:

1. The weak connections with a weight less than threshold pre-determined or non-connected pairs are selected from the frame before the prediction frame. The non-connected pairs that didn't meet with any events during the network evolution were eliminated.
2. The proximity scores of five different predictors (CN, JC, PA, AA, and our method) are computed for the chosen pairs.
3. The ranked top n -pair are taken into account as the future links or strong links that are more likely to appear.
4. If the weight of a pair with a weight less than *threshold* per-determined measure changes more than the even changing rate (e) from the frame under consideration to the prediction frame, the connection of the pair turn from weak to strong and vice versa.
5. The performance measures of all the predictors considered are compared.

5.8 Experimental Results

In order to evaluate the performance of the proposed method we conducted a set of experiments using DBLP-Citation-Network-V7 dataset. We downloaded the dataset from Arnetminer (<https://aminer.org>) [34]. The dataset contains 2,244,021 papers and 4,354,534 citation relationships. We used the validation set to empirically evaluate and determine the most appropriate values of p , i , r and α parameters in the link prediction task. We showed with some tests that the best performance was achieved at $p = 0.40, i = 0.80, r = -0.37$ and $\alpha = 0.1$. These results indicate that protective (p) and regressive (r) values almost balance with each other. The best performance values were achieved when $p < i$.

The first experiment is dedicated to evaluate the precision values of the proposed method and the traditional proximity scores on the unweighted network with $\{[1997-2000], [2001-2004], [2005-2008], [2009-2012]\}$ frames. In this structure, the frame size is 4 and the snapshot at $[2009-2012]$ time interval is used as the prediction frame. Precision can be evaluated at different points in a ranked list of extracted citations. Mathematically, precision at rank n ($P@n$) is defined as the proportion of relevant citations and extracted citations.

$$P@n = \frac{\text{number of relevant citations with rank } n}{n} \quad (5.15)$$

Table 2 Precision values (%) of five methods on the weighted network

Method	P@10	P@50	P@100	P@500
CN	5	7	28	44
JC	14	17	33	46
PA	13	23	32	57
AA	59	63	68	73
Our method (e=20%)	75	82	86	89
Our method (e=40%)	72	79	82	85

The second experiment finds the precision values of the proposed method and the other predictors on the weighted network. This experiment also investigates the effect of the event changing rate e for the performance of the time frame based score on the weighted citation network. Table 2 reports the results of this experiment. As can be seen from Table 2, the best prediction performances are achieved by our link predictor over first weighted version of the network with $e=20\%$. Since the event changes of the links (protective, inventive or regressive) between the frames are determined by this rate, it is important to choose the most appropriate event changing rate. This is because while the size of frame remains constant, if the event changing rate increases, less number of the event changes between frames is observed. The observations in Table 2 are similar to those described above for the frames on the unweighted network. However, it should be noted that all values are larger than those of the unweighted network

The third experiment deals with investigating the effect of the amortization factor α for the performance of the time frame based score. As mentioned before, this parameter emphasizes the importance of the second events in relation to primary events in the proposed method. For this purpose, we evaluated the time frame based score for four different values of α . It should be noted that when $\alpha = 0$, secondary events are not considered. However, for higher values, possible an extreme importance is assigned to secondary events. The AUC values obtained as measure of performance for different α values are given in Table 3. Our method performs best with an AUC value of 0.95 at $\alpha = 0.1$ on the weighted network with $e=20\%$. Second ranked are again the weighted network having $e=20\%$ with an AUC of 0.91 at $\alpha = 0.2$.

Table 3 AUC values of the proposed method for different values of α

Network	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Unweighted	0.68	0.82	0.92	0.87
Weighted ($e=20\%$)	0.72	0.86	0.95	0.91
Weighted ($e=40\%$)	0.70	0.83	0.89	0.88

6. CONCLUSION

There have been numerous attempts to address the problem of link prediction using supervised learning methods which acquire the knowledge through various measures. However, the knowledge gained through the existing static knowledge extraction measures are not sufficient for accurate link prediction in highly dynamic social networks. It leads to in accurate and less precise predictions. In order to alleviate this problem, we contributed by introducing two novel time frame based features which are based on common neighbors.

We focused on a network which indicate users information up to the year $T = 2012$ with a 4 year length of prediction window (i.e., new links prediction from 2009 up to 2012 is the task). We extracted $k = 3$ frames from the network structure and N obtained as $(N=F1, F2, F3)$.

In this thesis we introduced three novel temporal events namely (Protective, Inventive and Regressive) and the weighted version (W-Protective, W-Inventive and W-Regressive) which takes temporal information into account by monitoring how the network evolved along time. Our method consists of computing scores by aggregating rewards of the temporal events observed at each step in the network evolution, which resulted in more representative scores to the under analysis links.

Different from previous studies in this thesis we applied four topological proximity metrics (Common Neighbor index, Jacquard's Coefficient index, Preferential Attachment index and Adamic Adar index) both weighted and unweighted versions to the link prediction problem in a real citation network which is directed and weighted. Also the application of temporal information has considered in this study. Furthermore, we did a comparison between weighted and un-weighted metrics as the experiments show; the weighted metrics outperform the un-weighted metrics.

In the preformed experiments, we did a comparison between weighted and unweighted proximity metrics, according to the experimental results we can highlight that weighted proximity metrics outperform the unweighted metrics.

Experimental results applied on real dataset shows that the proposed method gives accurate prediction and promising results which outperform the traditional proximity metrics.

As we mentioned above all of the preformed experiments are according to the common neighbor based metrics and we didn't consider the global information based metrics and path

based metrics, therefore the application of mentioned metrics can be consider as future works, moreover this work still has limits regarding the case study, which the experiments deployed on one dataset and as future studies this work can extend by applying on different datasets.

REFERENCES

1. Wasserman, S., Faust, K: Social network analysis: Methods and applications (structural analysis in the social sciences). Cambridge University Press (1994)
- 2.
3. Getoor, L., Diehl, C. P.: Link mining: A survey. SIGKDD Explorations Newsletter, 7(2), 3–12 (2005)
4. Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In Proceedings of the 2007 seventh IEEE international conference on data mining pp. 322–331 (2007).
5. Wang P., Xû B., Wu Y., Zhou X.: Link Prediction in Social Networks: the State-of-the-Art. Sci China Inf Sci, 58: 011101(38), 2015.
6. Barabási, A. L., Jeong H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, 311(3-4):590 – 614 (2002)
7. Newman, M. E. J.: Clustering and preferential attachment in growing networks, April (2001)
8. Adamic, L. A., Adar E.: Friends and neighbors on the web. Social Networks, 25(3):211–230 (2003)
9. Salton, G., McGill M., J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA, (1986)
10. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol., 58:1019–1031, May (2007)
11. Lü, L., Zhou, T.: Link prediction in weighted networks: The role of weak ties. EPL (Europhysics Letters), 89(1):18001, (2010)
12. Murata, T., Moriyasu, S.: Link prediction based on structural properties of online social networks. New Generation Computing, 26(3):245–257, (2008)
13. Barabasi, A. L., Bonabeau E.: Scale-free networks, Scientific American, vol. 288, no. 5, pp. 60–69, (2003)
14. Zhou, T., Lü, L., Zhang, Y., C.: Predicting missing links via local information. Eur. Phys. J. B, 71:4 623-630, (2009)

15. De Sa, H., Prudencio, R., R., B., C.: Supervised Link Prediction in Weighted Networks. International Joint Conference on Neural Networks. 2281-2288, (2011)
16. Lin , Z., Xiong, L., Zhu, Y.: Link prediction using BenefitRanks in weighted networks. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. (2012)
17. Radicchi, F., Fortunato, S., Makines, B., Vespignani, A.: Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E.*, 80(5), 056103, (2009)
18. Tylenda, T., Angelova, R., Bedathur, S.: Towards time-aware link prediction in evolving social networks. In Proceedings of the third workshop on social network mining and analysis, SNA-KDD '09(pp. 1–9), (2009)
19. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A: Learning and predicting the evolution of social networks. *Intelligent Systems IEEE*, 25(4), 26–35, (2010)
20. Juszczyszyn, K., Musial, K., Budka, M.: Link prediction based on subgraph evolution in dynamic social networks, In Privacy security risk and trust (PASSAR) 2011 IEEE third international conference on social computing (pp. 27–34). (2011)
21. Jiang, C., Coenen, F., Zito, M.: A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(1), 75–105, (2013)
22. Huang, Z., Lin, D., K. J: The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2), 286–303. (2009)
23. Potgieter, A., April, K. A., Cooke, R. J. E., Osunmakinde, I. O: Temporality in link prediction: Understanding social complexity. *Journal of Emergence: Complexity and Organization*, 11(1), 83–96, (2009)
24. Soares, P., R., S., Prudencio R., B., C.: Proximity measures for link prediction based on temporal events. *Expert Systems with Applications*, 40(16), pp.6652-6660, (2013)
25. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In Proceedings of the 2003 international conference on information and knowledge management pp. 556–559, (2003)
26. Lü, L., Zhou, T. Link prediction in complex networks: A survey. *Physica A*, 390(6), 1150–1170, (2011)
27. Kaya, B., Poyraz, M.: Age-series based link prediction in evolving disease networks. *Computers in Biology and Medicine*, Vol: 63, 1-10, (2015)

28. Kaya, B., Poyraz, M.: Unsupervised link prediction in evolving abnormal medical parameter networks. *International Journal of Machine Learning and Cybernetics*, 7:1, 145-155, (2016)
29. Xiang, E., W.: A survey on link prediction models for social network data. *Science and Technology*, (2008)
30. Huang Z.: Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In *Proceedings of 12th ACM SIGKDD international conference on knowledge discovery and data mining*, (2006)
31. Kaya, B., Poyraz, M.: Supervised link prediction in symptom networks with evolving case. *Measurement*, 56, pp.231-238, October (2014)
32. Kaya, B., Poyraz, M.: Finding Relations Between Diseases by Age-Series Based Supervised Link Prediction. *ASONAM 2015*: 1097-1103, (2015)
33. Tan, P., N., Steinbach, M., Kumar, V., *Introduction to Data Mining*. (Vol. 1). Boston: Pearson Addison Wesley, (2006)
34. Schall, D.: Link prediction in directed social networks, *Soc. Netw. Anal. Min.* 4:157 (2014).
35. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*. pp.990-998, (2008)
36. Ryan Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. New perspective sand methods in link prediction. In *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 243–252, 2010.
37. Shou-De Lin and Hans Chalupsky. Unsupervised link discovery in multi relational data viararity analysis. In *Proceedings of 3rd IEEE International Conference on Data Mining*, pages 171–178, 2003.
38. Sotiris B. Kotsiantis, Ioannis D. Zaharakis, and Panayiotis E. Pintelas. Machine learning: are view of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.
39. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3), 1995.
40. Janardhan Rao Doppa, Jun Yu, Prasad Tadepalli, and Lise Getoor. Learning algorithms for link prediction based on chance constraints. In *Proceedings of European Conference Machine Learning and Knowledge Discovery in Databases*, pages 344–360, 2010.

41. Myunghwan Kim and Jure Leskovec. Modeling social networks with node attributes using the multiplicative attribute graph model. In Proceedings of 27th Conference on Uncertainty in Artificial Intelligence, pages 400–409, 2011.
42. Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit S. Dhillon. Supervised link prediction using multiple sources. In Proceedings of 10th IEEE International Conference on Data Mining, pages 923–928, 2010.
43. Milen Pavlov and Ryutaro Ichise. Finding experts by link prediction in co-authorship networks. In Proceedings of 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics, pages 42–55, 2007.
44. Horvath, S., Weighted Network Analysis. Applications in Genomics and Systems Biology. Springer Book. ISBN 978-1-4419-8818-8. 2011.
45. Granovetter, M., The strength of weak ties. American Journal of Sociology 78(6), 1360-1380. 1973.
46. Kanna Al Falahi, Yacine Atif, Said Elnaffar, Social Networks Challenges and New Opportunities, IEEE/ACM International Conference on Cyber, Physical and Social Computing, pp: 804-808, 2010.
47. P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison Wesley, 2287. 2005.
48. T. Zhou, L. U. L. and Y. C. Zhang. Predicting missing links via local information. Eur. Phys. J. B, 71, 623, (2009).
49. Discrete Mathematics and its Applications, Kenneth H. Rosen, 7th Edition

CURRICULUM VITA

Mujtaba Jawed

Mujtabajawed786@gmail.com

Nationality: Afghan
Place of birth: Kabul
Date of birth: 25 / 7 / 1989
Marital status: Single

EDUCATION

09/2013 – 03/2016 MCE (Master of Computer Engineering), Firat University, Elazig, Turkey.
09/2014 – 02/2015 MCE (Master of Computer Engineering), Lodz University, Lodz, Poland.
10/2012 – 07/2013 Certified From Gaziantep University TOMER (Turkish Language), Turkey.
03/2006 – 12/2010 Bachelor of Computer Science, Kabul Education University, Afghanistan.
2006 High School, Muhammad Alam Fizard high School Kabul, Afghanistan.
2004-2005 Graduated from Hewad English Language Center, Kabul, Afghanistan.

WORK EXPERIENCES

10/11/2011 – 04/11/2012

- Worked as an IT Specialist in Black & Veatch USAID in Kandahar province.

06/07/2011 – 02/11/2011

- Worked with DABS (Da Afghanistan Breshna Sherkat) as a IT Technician in Kandahar province

02/07 /2010 – 06/06/2011

- Worked in Maah Publication as IT Officer and Text manager.

02/08/2009 – 29/06/2010

- Worked in Ministry of Information and communication Technology as an IT officer.

PUBLICATIONS

1. Jawed, M., Kaya, M., Alhajj, R.,: Time Frame based Link Prediction in Directed Citation Networks. ASONAM 2015: 1162-1168, (2015)
2. Jawed, M., Kaya, M., Butun, E,. Alhajj, R Unsupervised Link Prediction based on time frames in Weighted-Directed Citation Networks. Lecture notes in social networks: (2016)