University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 4, Issue 2, July 2020
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

A survey of structural node-similarity methods for link prediction in complex networks

Bisharat Rasool Memon[1], Kamran Dahri[1], Abdul Waheed Mahesar[1], Zia Ahmed Shaikh[2]

[1] Department of Information Technology, University of Sindh, Jamshoro, Pakistan
[2] Directorate of Information Technology, Liaquat University of Medical and Health Sciences, Jamshoro, Pakistan
bisharat.memon@usindh.edu.pk, kamran.dahri@usindh.edu.pk, waheed.mahesar@usindh.edu.pk, zia.shaikh@lumhs.edu.pk

Abstract: This paper surveys several commonly used techniques for link prediction in networked/relational systems. This survey considers the body of literature from networks science, social networks analysis, and related research, and surveys several well-known analytical methods based on structural similarity of participating nodes. These methods that have been or could be used for solving the problem of link prediction in networked systems. The paper starts with a formalization of the link prediction problem previously given in the context of social networks. We discuss the notion of structural similarity among nodes in a network, and why and how these structure-derived node similarity measures also quantify the likelihood of the presence of future links in the network. The surveyed methods include proximity indices based on graph-theoretic distances between nodes, as well as, on local and global neighbourhoods. The authors identify and discuss a number of challenges which complicate link prediction due to certain conditions, or due to the necessity of consideration of exogenous factors to the network rather than just its endogenous structural properties.

Keywords: complex networks; social networks; link prediction; similarity indices; network topology; network structure

I. INTRODUCTION

Many real-life systems generate "relational data", i.e., data having properties (of interaction) in addition to the descriptive attributes of the entities involved in the system. Such systems are often best modelled as graphs or networks to capture the regular pattern of relations among the constituent entities—that is, the *structure* of the system. Networks are composed of nodes and links to represent respectively the entities and relationships (among those entities) in some system.

Prediction is an important problem in data mining. In the context of complex networks, it is and interesting to understand which links will appear and which links will disappear in the future. This is a natural consequence of the dynamic nature of complex networks—they change (grow or shrink) not only in their size but also in their structure—forming new links or breaking existing links among the nodes of the network.

A number of factors influence the dynamics of link formation in the network [1], [2]. One factor is the attributes associated with both node and link type entities in the network. On the one hand these can be *compositional* attributes (e.g., age, gender, role, bandwidth, distance, etc.), which describe the entity that the node or the link represents. These are exogenous to the network itself. On the other hand, the actual topology of the network gives rise to a number of structural properties of the nodes, links (e.g. node and link centralities, etc.), and of the network as a whole. These attributes are derived purely from the particular pattern of connections that the network may have at any given time, and are therefore, liable to change overtime as the network structure changes. Thus, the descriptive (or compositional) and structural attributes of nodes, in their own right as well as together, play a substantive role in determining the structure of the network at any time.

In case of most social networks, it is the nodes' descriptive attributes (such as, age, gender, geography, race, shared associations, etc.) that determine the particular pattern of the instances of one or more social relationships in that network. Although such attributes may be responsible for the initial structure when the network is beginning to take shape, the formation of future links is also influenced by the endogenous attributes of the nodes such as their positions in the network [3]. This leads to the important realisation that the future structure of a network (i.e., the particular pattern of links) can be derived from the network's own structural properties which are latent in its topology at any given time. This reasoning underlies all the node-similarity techniques surveyed in this paper, and which form the basis for the respective link prediction technique.

The rest of the paper is structured as follows: Section II gives a formal description of the *Link Prediction Problem (LPP)* in complex networks. It describes the documented generic experimental set-up for all the prediction approaches discussed in this report, and the method for evaluating their effectiveness. Section III elaborates upon the idea of node similarity in the structural context and describes why it is a valid notion upon which link prediction can be based. Section

IV discusses how link formation likelihood is a function of the graph-distance between nodes. Sections V and VI discuss individual node similarity indices based on local and global structural characteristics, respectively. Section VII identifies some challenges that complicate link prediction under certain conditions, or when it is important to consider other factors in addition to just network structure. Section VIII briefly mentions other network analysis problems that are similar or closely related to link prediction, and also mentions several practical applications of link prediction in other fields. Section IX concludes this paper by summarising key points and provides some pointers to other classes of link prediction approaches not discussed herein.

II. The Link Prediction Problem

The problem of predicting links in social networks has been formalized by Liben-Nowell and Kleinberg [4] as the *Link Prediction Problem* (LPP). The formalisation can easily be extended to networks in general:

> Given a snapshot of a network at some time t, ... [the objective is] to accurately predict the edges that will be added to the network during the time interval from time t to a given future time t'. [4]

A. Problem Description

A network is usually represented as a graph $G = (V, E)$, where V is the vertex set, and E is the edge set. If the entities represented by vertices u and v are directly connected under some relationship, then the edge $(u, v) \in E$, otherwise $(u, v) \notin E$. It is assumed that the network is evolving, so new edges are formed, and the network structure keeps changing with time. Each edge $e \in E$ has a timestamp $t(e)$ on it to keep track of this evolution. $G[t, t']$ is the sub-graph G consisting of all edges between times t and t', i.e., $G[t, t'] = (V',E')$, $V' = \{ x \mid x \in \{u, v\} \exists (u, v) \in V' \}$, $E' = \{ e' \mid t \leq t(e') \leq t' \}$

By choosing four different times $t_0 < t'_0 < t_1 < t'_1$, the original graph can be broken into two sub-graphs: $G[t_0, t'_0] = (V, E_0)$ and $G[t_1, t'_1] = (V, E_1)$, akin to having *observed* and *unobserved* sub-networks, respectively. The former provides the data for training and the later for testing. Each different approach to measuring node similarity can use the structural information from the observed sub-graph to predict edges in the unobserved sub-graph. For the sake of simplicity, it can be assumed that the vertices in both these sub-graphs are the same, i.e., no new vertices were added to or removed from the network between times t_0 and t'_1. So, any solution to the link prediction problem should output a list of edges not present in $G[t_0, t'_0]$ but are likely to appear in $G[t_1, t'_1]$. Therefore, the set of new edges to be predicted is $E_{new} = E_1 - E_0$. Based on the particular node similarity approach, each prediction method assigns a likelihood score $score(x, y)$ to each unobserved edge $(x, y) \in E_{new}$. The likelihood values for $score(x, y)$ are particular to the methods for measuring node similarity and scores from different approaches cannot be compared. However, the performance of different prediction approaches can be evaluated using a uniform heuristic based on the size of the sets of true positives (described in the next section).

B. Evaluation Effectiveness of Prediction

A list of all edges $(x, y) \in E_{new}$ is ranked by the likelihood $score(x, y)$ and for some positive k, the top k edges are picked, as the set E_{new}^k. A measure of accuracy of the predictor is the size of intersection of the set E_{new}^k with the set of all the edges actually present in the given network, i.e., E_1 (the set of all the edges that can ever be present, $V \times V$ can also be used.)

$$Predictorofeffectiveness = |E_{new}^k \cap E_1|, \text{ or}$$
$$Predictorofeffectiveness = |E_{new}^k \cap (V \times V)|$$

III. Node Proximity or "Similarity"

The often-expressed adage "birds of a feather flock together" is more than just a cliche. In many real-life situations it is true that similarity breeds association, and this is more true as an observed phenomenon in social networks than anywhere else [5]. People tend to form connections with others based on shared attributes, such as, age, tastes, beliefs, interests, political leanings, class, organisational roles, etc. This tendency among individuals to associate or bond with others who are perceived to be similar is called *homophily*, literally meaning "love of the same".

In a network, node similarity is based on some attributes of the nodes. The extent to which two or more nodes are similar depends on the extent to which they have common attributes. These attributes of nodes can be either structural or compositional attributes. However, in this work we discuss indices of node "similarity" or "proximity" based purely on their structural or topological attributes, which are intrinsic to the network structure. Some approaches of measuring node similarity are based on nodes' specific compositional attributes (not all compositional attributes have the same importance to similarity), which are exogenous to the network structure, and can vary from one domain to another. All the so-called *similarity-based indices* depend on methods that assign a connection weight $score(x, y)$ to pairs of nodes x and y, based on the input graph, and then produce a ranked list in decreasing order of $score(x, y)$. This can be viewed as computing a measure of structural "proximity" or "similarity" between nodes x and y, which gives a prediction likelihood for any future link between the two nodes. Node similarity can either based on *local* position of nodes determined by their immediate neighbourhood, or by the overall *global* position determined by the overall structure of the network. Indices based on both these notions of node similarity are discussed under separate subheadings.

IV. Shortest-path Distance as a Predictor of Future Links

Milgram [6] showed that real-world social networks are characterised by the small world phenomenon, where any two people in the world are connected through a short chain of acquaintances. Many other real-life network phenomena exhibit "small-world" properties [7]. Several studies have shown that social networks, the World Wide Web, gene networks, neural networks, power grids, road networks, all exhibit small-world network characteristics. Small-world networks are characterised by small average geodesics

(shortest path distances), and large clustering coefficients. The former characteristic means that most nodes of the network are separated by small chains of not more than a few links. In the context of social networks, it means that it is often possible to link total strangers through a mutual acquaintance—the so-called *small-world phenomenon*. Studies on scientific collaboration networks have shown that such networks are *small worlds*, in which randomly chosen pairs of scientists are separated by a chain of a few intermediate acquaintances [8], [9]. It follows from this notion, that individuals connected by fewer intermediate links are more likely to form direct connections. In the above example, if two non-collaborating scientists have one or more common collaborators (represented by two nodes having a non-empty common neighbourhood), it would increase the future likelihood of collaborating themselves. Based on this reasoning, the shortest-path distance can be considered as a very crude measure of link-prediction. In this case, the prediction likelihood, *score(x, y)* of a future link between nodes *x* and *y* is the negative of the shortest path distance between them:

$$score(x, y) = -dist(x, y) \quad (1)$$

V. Local Similarity: Methods Based on Node Neighbourhoods

Several methods of measuring structural similarity between two nodes are based on node neighbourhoods. The idea is that two random non-adjacent nodes are more likely to form a link in the future if there is a large enough overlap between their neighbourhoods, i.e., there are many other nodes as common neighbours. Several indices based on the idea of common node-neighbourhoods (originally developed to serve different purposes) can be adapted as predictors to give likelihood scores for future links. The intuition comes from a social (friendship or collaboration) network, where two individuals who share many of the same friends or collaborators are more likely to come into contact themselves and form a direct social link. Jin et al. [10] have proposed models of social network evolution based on the principle that the edge *(x, y)* is more likely to come into existence if for some *z* the edges *(x, z)* and *(y, z)* already exist.

If *x* is a node in the network *G(V, E)*, we denote the set of all nodes in the immediate neighbourhood of *x* by *Γ(x)*, such that:

$$\Gamma(x) = \{y : (x, y) \in E\} \quad (2)$$

A. Common Neighbours

The simplest approach in this category of prediction scorers is to simply take the size of the overlap between neighbourhoods of two nodes *x* and *y* as the likelihood score of a future link between the two nodes. In a social network context, it has been verified that there is a correlation between the size of neighbourhood overlap for two nodes *x* and *y* and the probability of them forming a direct social connection in the future [11]. Thus, we have:

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (3)$$

The above measure is not affected by the size of node neighbourhoods only by how much those neighbourhoods overlap, which can give rise to some problems of comparison.

In order to normalise the above measure to ease comparison across different networks, we can divide the score by *(n-1)* as the maximum number of adjacencies any node in the network can have; we have, $score(x, y) = |\Gamma(x) \cap \Gamma(y)| / (n-1)$.

B. Jaccard's Index

Jaccard's index or coefficient of similarity (also synonymously, Tanimoto Similarity [12]), originally proposed as a statistic for measuring similarity between two finite sample sets over some features that either or both sets could have. The measure of similarity is given as a ratio of the number of features common to both sets to the number of features either one or the other set has [13]. In the context of node neighbourhoods in a network, the "feature" we are interested in is common neighbours of nodes *x* and *y*. Jaccard's similarity coefficient, adapted as a measure of the likelihood of a future link between the two nodes, is given as:

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (4)$$

C. Adamic/Adar Similarity

The study of individuals' homepage networks by Adamic and Adar [14], shows that certain structural features (in and out links between pages) and other exogenous features (such as, mailing list subscription, interests mentioned in text, etc.) are reflections of social interactions the users have in the real world. This study proposes a node similarity metric to predict existence of link between pair of nodes in a social network, based on counting the number of similar features both nodes have—giving more weight to shared features that are rarer, and less weight to features that are more common. This metric is expressed as:

$$similarity(A, B) = \sum_{shared_{i}tems} \frac{1}{\log[freq(shared_{i}tem)]} \quad (5)$$

The above equation can be adapted for link prediction based on common neighbourhoods. If *z* is a common neighbour of nodes *x* and *y*, then taking *z* as a shared feature, we have the following equation for our link prediction likelihood score:

$$similarity(A, B) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log[|\Gamma(z)|]} \quad (6)$$

The inverse log frequency in the above quantity penalises common neighbours (*z*) that are themselves linked (choose / chosen by) a lot others, and rewards when they are more exclusive to the pair.

D. Preferntial Attachment and the Rich Club Effect

Preferential attachment is a model of network growth proposed by Barabasi and Albert [15] based on the notion of rich-get-richer. The basic idea is that new nodes joining the network tend to attach themselves to nodes with high degrees, i.e., nodes which already popular and have many other direct connections. The resulting network is scale-free in that the degree distribution follows a power law. If x is a node in a network that grows according to preferential attachment, then the probability of x forming a new link within the network is proportional to the size of its neighbourhood, i.e., $|\Gamma(x)|$. What should then be the probability of node x forming a link with some other node y? According to empirical evidence from studies on scientific collaboration networks [16], [17], the probability of link (x, y) forming is correlated with the product of the neighbourhood sizes of x and y. Same result is reported for scale-free networks without growth [18].

Additionally, real-world networks (especially, those involving individuals and organisations) often exhibit the so-called *rich club effect* [19], [20] with respect to the number of incoming ties viewed as a resource for nodes in the network—nodes with higher degrees (many neighbours) tend to be more inter-connected than nodes with lower degrees (few neighbours).

This reasoning gives the corresponding similarity measure as follows:

$$score(x, y) = |\Gamma(x)|.|\Gamma(y)| \tag{7}$$

VI. Global Similarity: Methods based on Set of all Paths

In this section we discuss several methods that extend the idea of node similarity based on graph-distance by considering all the paths (not just the shortest path) between two nodes. Algorithms for global similarity indices are more expensive than those for local similarity indices, because much more information about the structure of the network is required.

A. Katz' Index

Katz [21] proposed a method for measuring node status (centrality) that takes into account for a given node, not only *how many* choices are received but also *who* makes those choices. The resulting index measures node status by counting all the paths of length l between nodes x and y, dampened exponentially by length to give shorter paths more weight. For the nodes x and y we have:

$$socre(x, y) = \sum_{l=1}^{\infty} \beta^l (A^l)_{xy} = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots \tag{8}$$

where β is the damping factor whose value should be less than the greatest eigenvalue of the adjacency matrix A for the series to converge. The above measure gives the corresponding xy entry in the closed-form matrix of Katz' score for all nodes, i.e., $(I - \beta A)^{-1} - I$.

B. Hitting Time and Commute Time

Another indicator of distance (inversely, proximity) between two nodes is the expected time (as number of steps) it takes a random walker starting from one node to reach the other node, known as *hitting time*, $H_{x,y}$. In general hitting time is not symmetric, i.e., $H_{x,y} \neq H_{y,x}$, and it is more natural to consider *commute time*, $C_{xy} = H_{x,y} + H_{y,x}$, which is the expected number of steps for a random walker to arrive from x to y and back. Both *hitting time* and *commute time* (negated) can serve as natural measures of node proximity, and hence, similarity.

$$score(x, y) = -H_{x,y} \tag{9}$$

$$score(x, y) = -\left(H_{x,y} + H_{y,x}\right) \tag{10}$$

In general, if x and y are nodes in a connected networks, the hitting time $H_{x,y}$ is given by:

$$H_{x,y} = \begin{cases} 0, x = y \\ 1 + \frac{1}{\Gamma(x)} \sum_{z \in \Gamma(x)} H_{x,y}, x \neq y \end{cases} \tag{11}$$

where $z \in \Gamma(x)$ is a neighbour of x.

A problem that negatively affects hitting time and its variants is their sensitivity to parts of the graph distant from the source and destination nodes (i.e., topological noise) even when the two nodes are connected by short paths. This difficulty can be overcome by adapting the *rooted PageRank* as a random walk with "restart".

C. PageRank and Random Walk with Restart

PageRank is a very well-known algorithm for ranking webpages on the Web based on their "importance" and forms a basis for Google's Web search algorithms [22]. Under PageRank, the importance and relevance of a webpage is correlated to the number of links it receives from other webpages. In addition to the Web graph (with webpages and hyper-links as vertices and edges, respectively), it is also well defined for any graph in general.

PageRank uses a positive real value, $\alpha \in [0,1)$, to control the "diffusion" of the random walk and allowing for periodically resetting of the random walker to restart from the source. To obtain a metric for node similarity $score(x, y)$, the rooted PageRank can be adapted by considering a random walker starting from x, that iteratively moves to a random neighbour of x with a fixed probability $\alpha \in [0,1)$ and returns to x with probability $1 - \alpha$. If $q_{x,y}$ is the probability that the random walker starting from x locates y, then $q_{x,y}$ is the y^{th} element in the vector q_x, given as [23]:

$$\overrightarrow{q_x} = (1 - \alpha)(I - \alpha P^T)^{-1} \overrightarrow{e_x}, \tag{12}$$

where P is the transition probability matrix with $P_{xy} = 1/deg(x)$ if x and y are connected, otherwise 0; and e_x is a vector whose x^{th} element is 1 and others 0.

The node similarity score $score(x, y)$ based on random walk with restart is then defined as:

$$score(x, y) = q_{xy} + q_{yx} \quad (13)$$

D. SimRank

SimRank follows from the intuition that objects that are similar are related to other objects that are themselves similar. According to this recursive notion of similarity, two nodes are similar to the extent that they are connected to (other) similar nodes, with the base case that any node is completely similar to itself [24]. For example, nodes *i* and *j* are similar if they are connected to similar nodes *m* and *n* respectively. Mathematically, we have:

$$similarity_{xy} = \begin{cases} 1, & x = y \\ \gamma . \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} similarity_{ab}}{|\Gamma(x)|.|\Gamma(y)|}, & x \neq y \end{cases} \quad (14)$$

where $\gamma \in [0, 1]$ is the decay factor.

VII. Some Challenges

The approaches to link prediction discussed in this report all depend on only the intrinsic characteristics of the observed network, and do not consider a number of other important factors that complicate the task of link prediction.

Although it is true that problems from many diverse domains can be mapped to networks, they don't always follow the same abstract model of network evolution, and the likelihood of link formation is as well a function of node and link attributes extrinsic to network, as it is of structural characteristics intrinsic to the network. The challenge is, therefore, to combine node and link attributes with topology so that the resulting network evolution models are more realistic and applicable. Interpretation of node similarity differs with domain context is there is no standard approach of prediction that combines information encoded in nodes and links with topological information.

Another issue that complicates link prediction is network sparsity. Real life networks are often sparse, that is, the number of links actually present is a very small proportion of all the possible links that can be—in short, the network density is very small. This is often not because there are so few interactions, but because the number of nodes is too great. Consequently, achieving good link prediction accuracy for sparse networks is a big challenge, where in a network of size on the order of millions of nodes (with billions of "possible" connection) the chances for always getting false positives are impossibly high. Indeed, a predictor can be very accurate by predicting no links at all.

Yet another challenge in link prediction comes from heterogeneity in networks. Many real-world systems, modelled as complex networks, have multiple interactions among entities of multiple types—i.e., heterogeneity in nodes as well as links. "Link prediction in such networks must model the influences between heterogeneous relationships and distinguish the formation mechanisms of each link type, a task which is beyond the simple topological features commonly used to score potential links." [25] So far, little work has been done in link prediction in heterogeneous networks. Some recent related work in this area is done by Davis et al. [25], [26].

VIII. Related Problems and Applications

The natural applicability of networks in a variety of fields means that any environment that maps to a network usually has a well-defined important question in that environment that maps back to the link prediction problem in networks. This section briefly mentions some typical applications on link prediction.

A link may relate more than two entities—the equivalent graph-theoretic concept is of hyper-edges which can have more than two endpoints. A prediction related problem is *link completion* where the link is known or observable, but it is incomplete. Given one or more nodes that are known to have a link, the objective is to determine which other nodes are also endpoints of that link.

Another problem in network analysis that can viewed as a generalisation of link prediction is the *network completion problem*, where the observed network is missing links as well as nodes, and the objective is to accurately predict all the missing elements of the network. This scenario is encountered in many real-life situations where it is either impossible to know which elements belong to the network, or the collected data is incomplete with respect to both the nodes and the links. Kim and Leskovec use a scalable approach based on an expectation maximisation model to complete a network where as many as half the nodes are missing [27].

Sometimes a network may have false links due to either unintentional errors in data collection or deliberate misreporting to mislead investigators and prevent discovery of the real network structure. Link prediction methods can be used to identify these spurious links in addition to predicting any missing links. However, when using link prediction methods to remove spurious link, it is important to consider that there might be some real but "unexpected" links.

Recommender systems use link prediction to make recommendations, suggest relevant products or services, recommend friendships in online social sites like LinkedIn, Facebook, twitter, etc. Li and Chen study recommendation as a link prediction problem in bipartite graphs and use a machine learning approach based on a combination of nodes' descriptive features (demographics, etc.) and graph-based structural features [28].

In a covert-network context, e.g., monitoring of a criminal or terrorist network, link prediction allows investigators to make conjectures about possible connection between individuals whose interactions have hitherto gone unobserved [29]. Moreover, in countering hostile networks, one of the objectives is to neutralise one or more critical nodes to destabilise the network. However, sometimes networks can be very resilient, and can recover from destabilisation attacks [30], by forming new links to restructure and regroup. It then becomes important to test several alternative hypotheses of node removal to determine the least costly post-removal scenario.

IX. CONCLUSION

Link prediction in complex networks is a very important research issue, not only from a purely network analysis perspective, but also because it has equivalent practical applications in nearly all of the many systems which can naturally be modelled as networks. The formation of links between a pair of nodes is a function of many factors, including descriptive and structural attributes of the participating nodes. In this report, several typical methods of link prediction based on measures of topographical node similarity were discussed in detail. However, it is not an exhaustive review of all the link prediction methods. Approaches other than those based purely on node similarity algorithms include methods based on network evolution models, where prediction is based on testing against certain properties suggested by the model—e.g., random graph, Barabasi-Albert [15], Watts-Strogatz [7], etc. A number of challenges that complicate link prediction in networks were also highlighted. Some of these challenges offer directions for future research. For example, link prediction in directed and heterogeneous networks, can be important areas for future work in this regard, complimenting previous work with weighted [31] and heterogeneous [32] networks.

REFERENCES

[1] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010.

[2] M. E. J. Newman, Networks: An Introduction. OUP Oxford, 2010.

[3] H. C. White, S. A. Boorman, and R. L. Breiger, "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions," American Journal of Sociology, vol. 81, no. 4, pp. pp.730–780, 1976.

[4] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," Journal of the American Society for Information Science and Technology, vol. 58, no. 7, pp. 1019–1031, 2007.

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," Annual Review of Sociology, vol. 27, no. 1, pp. 415–444, 2001.

[6] S. Milgram, "The Small World Problem," Psychology Today, vol. 1, p. 61, 1967.

[7] D. J. Watts, "Networks, Dynamics, and the Small-World Phenomenon," American Journal of Sociology, vol. 105, no. 2, pp. 493–527, #sep# 1999.

[8] M. E. J. Newman, "The Structure of Scientific Collaboration Networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no. 2, pp. pp.404–409, 2001.

[9] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," Physical Review E, vol. 64, no. 1, pp. 1–7, #jun# 2001.

[10] E. M. Jin, G. Michelle, and M. E. J. Newman, "Structure of Growing Social Networks," Physical Review E, vol. 64, p. 046132, 2001.

[11] M. E. J. Newman, "Analysis of weighted networks," Physical Review E, vol. 70, no. 5, p. 9, #jul# 2004.

[12] D. J. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants," Science, vol. 132, no. 3434, pp. 1115–1118, 1960.

[13] P. Jaccard, "The Distribution of the Flora in the Alpine Zone," New Phytologist, vol. 11 (2), pp. 37—50, 1912.

[14] L. A. Adamic and E. Adar, "Friends and Neighbors on the Web," Social Networks, vol. 25, no. 3, pp. 211–230, 2003.

[15] A.-L. Barab´asi and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, no. 5439, pp. 509–512, 1999.

[16] A.-L. Barab´asi, H. Jeong, Z. N´eda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the Social Network of Scientific Collaborations," Physica A: Statistical Mechanics and its Applications, vol. 311, no. 3–4, pp. 590–614, 2002.

[17] M. E. J. Newman, "Clustering and preferential attachment in growing networks," Phys. Rev. E, vol. 64, p. 025102, #jul# 2001.

[18] Y.-B. Xie, T. Zhou, and B.-H. Wang, "Scale-free Networks without Growth," Physica A: Statistical Mechanics and its Applications, vol. 387, no. 7, pp. 1683–1688, 2008.

[19] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," Nature Physics, vol. 2, pp. 110—115, 2006.

[20] M. E. J. Newman, "Assortative Mixing in Networks," Phys. Rev. Lett., vol. 89, p. 208701, #oct# 2002.

[21] L. Katz, "A new status index derived from sociometric analysis," Psychometrika, vol. 18, no. 1, pp. 39–43, #mar# 1953.

[22] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," Computer Networks and fISDNg Systems, vol. 30, no. 1–7, pp. 107–117, 1998, proceedings of the Seventh International World Wide Web Conference.

[23] L. L¨u and T. Zhou, "Link prediction in complex networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, #mar# 2011.

[24] G. Jeh and J. Widom, "SimRank: A Measure of Structural-context Similarity," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 538–543.

[25] D. Davis, R. Lichtenwalter, and N. V. Chawla, "Multi-relational Link Prediction in Heterogeneous Information Networks," 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 281–288, #jul# 2011.

[26] D. Davis, R. N. Lichtenwalter, and N. V. Chawla, "Supervised methods for multi-relational link prediction," Social Network Analysis and Mining, vol. 3, no. 2, pp. 127–141, #apr# 2012.

[27] M. Kim and J. Leskovec, "The Network Completion Problem: Inferring Missing Nodes and Edges in Networks," in SIAM International Conference on Data Mining (SDM) 2011, 2011.

[28] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," Decision Support Systems, vol. 54, no. 2, pp. 880–890, #jan# 2013.

[29] V. Krebs, "Mapping Networks of Terrorist Cells," CONNECTIONS, vol. 24, no. 3, pp. 43–52, 2002.

[30] R. Lindelauf, P. Borm, and H. Hamers, "Understanding Terrorist Network Topologies and Their Resilience Against Disruption," in Counterterrorism and Open Source Intelligence, ser. Lecture Notes in Social Networks, U. K. Wiil, Ed. Springer Vienna, 2011, vol. 2, pp. 61–72.

[31] B. R. Memon, "Identifying Important Nodes in Weighted Covert Networks Using Generalized Centrality Measures," in 2012 European Intelligence and Security Informatics Conference. IEEE, #aug# 2012, pp. 131–140.

[32] B. R. Memon and U. K. Wiil, "Visual Analysis of Heterogeneous Networks," in 2013 European Intelligence and Security Informatics Conference, 2013, pp. 129–134.