Missing Link Prediction in Complex Networks

M. Usman Akhtar[1], Iftikhar Ahmad[1], M. Imran Khan Khalil[1], Sheeraz Ahmed[2]
[1] University of Engineering and Technology, Peshawar, Pakistan
[2] Iqra National University, Peshawar, Pakistan

***Abstract*— Real world complex networks are indirect representation of complex systems. They grow over time. These networks are fragmented and raucous in practice. The important concern about complex network is link prediction. The crux of link prediction is to determine the possibility of probable edges. The link prediction demand is very often spotted in recommendation systems, such as recommending new friends on social networks, or recommending shopping cart based on earlier online product associated searches and purchases. In this research we study link prediction approaches. We compare different link prediction setup in term of their prediction performance using various performance metrics, such as True Positive, False positive, and Area Under Curve (AUC). Our results on real world complex network data shows that most of openly identified attributes are very easy to quantify and surprisingly effective in link prediction problem. Common Neighbor and Distance outperforms other algorithm with narrow margin in variety of different performance measures. During link prediction a small batch of attributes always plays a remarkable role in the link prediction job.**

***Keywords*— Complex Networks, link Prediciton, Social Network Analysis, Graph Theory**

1. Introduction

Networks are effective descriptions of real world complex networks [1]. Complex networks describe the interaction among the elements of complex systems such as computer [2], neural, chemical [3] and online social networks [4]. In these networks, entities (such as computer, neurons, chemical agents etc) are represented by nodes (also called vertices), where edges between pair of nodes depicts interactions/associations between the nodes.

Complex networks have application in many branches of science [5]. It has been applied in epidemiology to predict disease and virus spreading in communities [6] and plan vaccination to try to inhibit it. Likewise the complex network analysis can be applied in political campaigns to influence maximum voters [7], and in transportation planning to improve routes [8]. A lot of efforts have been made to comprehend the network evolution [9, 10], and the fundamental topological structure of complex real world graph [11].

One crucial scientific issue related to complex network analysis is link prediction [12]. Networks are very agile in nature, fresh vertices and edges are added over the passage of time [13]. Basic idea of link prediction is to approximate the possibility of the existence of a link between pair of nodes, derive from perceived topological structural attributes of nodes [14]. For example, in online connected community networks, future associations can be suggested as likely-looking friendships, which can assist users in recommending new friends and thus strengthen their dependability to the service [10]. In other words, link prediction provides a measure of social appropinquity between pair of nodes. The only available information is topological structure of the network [15]. Applications of the phenomenon include suggestion of new followers/friends one social websites such as Google Plus, Facebook, Foursquare, LinkedIn, Twitter etc. In addition, it can also be used to suggest interests that are most likely collective. e.g. Amazon, Netflex and Google AdWords [16].

Against each network there are a number of algorithms, however they lack experimental validation for large set of data set. Algorithm tested on validated against 1 or two specific datasets, algorithm vs algorithm, we don’t know which is better.

2. Formal Problem Setting and Contributions

Assume a network $G(V,E)$ represent an undirected network at a particular time t, where V and E represents set of nodes and edges respectively. Self-connections and multiple-links are not allowed [17]. L is the total number of missing links in E. Each edge $e = (u,v) \in E$ depicts interaction between nodes u and v at time t. The link prediction aim is to predict links for approaching time $t'(t' > t)$ in existing network [14], $t + t'$

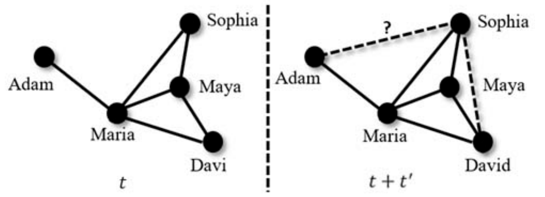

Figure 1 Definition of simple network, dashed line depicts possible link between non-adjacent nodes

This problem is clarified with simple network of about 5 persons in Fig 1, The total number of possible links in simple 5 node network$\frac{5(5-1)}{2} = 10$. For the missing link $U - E$, the prediction task is to know the fundamental mechanism of link formation in particular complex network and using the current topological structure properties to estimate the non-existing links probability. Solid line represents links in network at time t, and dashed line representing the link that may occur in future $[t, t']$, Maria and Adam are friends Maria and Sophia are also friends at time t'. Possibly Maria introduce Sophia with Adam, they become friend too. Similarly, Sophia and David may become friend at time t'.

Normally we are not aware which links may occur in future, otherwise we do not need link prediction [18]. The primary goal of link prediction problem is to forecast new link that may take place in near future. The discussed link prediction work is under assumption, may not hold true for each and every complex network. Apparently apprehending the ultimate growth device of social network is yet not easy because of diverse behavior difficult because of diverse behavior and characteristics of individual [2]. In this paper we compare similarity-based link prediction algorithms to infer the best algorithm out of them for real world complex networks.

Algorithms performance is argued on the ratio of correctly predicted links in probe set E^P. Experiment is repeated 5 times time for each dataset on four under consideration algorithms and their average result is computed multiple times. The network of future is overwhelmed by arrangement and fame of nodes. The sighting of links or collaborations is costly [19]. An exceptionally correct assumption can lessen the cost of experiment and improves the speed of unleashing the future.

Rest of the paper is organized as follows. Section 3 briefly present the literature review of related works in the field of missing link prediction. Section 4 presents the methodology adopted for this paper. The set of algorithms under consideration, and the data sets are also described in said section. While Section 5 posters results and assimilates discussion, whereas Section 6 based on conclusion of the research work.

3. Literature Review

Wang et al [2] presented an algorithm that made use of current popularity of node based on the idea that an active node have more probability of attraction to future nodes. In combination with a proposed innovative approach entitled popularity based structural perturbation method (PBSPM). PBSPM is a similarity based approach that measures the possibility of links through knowing collective aspects, i.e. common friends, age differences, professions, and tracing locations on which the two end points have in common. First of all, the network will be divided into two sets i.e. the training set E^T and the probe set E^P. Each edge has the bases on the birth time of each edge. Further elaboration of the training set E^T is bisected in to two parts, i.e. the old set and the new one in order to sum up the popularity, results are obtained by experimenting ten times every time E^T is obtained randomly, x and y are two nodes which are scored S_{xy} i.e. missing links according to their current popularity, and existing topological structure. The approach is affected due to unreachable as well as undependable information of nodes because of privacy policy in everyday life. And may only be valid for only few real world scenarios.

Yang and Zhang [20], introduced an algorithm based on common neighbors and distance to predict link in variety of real world networks from available topological structure of complex networks, without increasing complexity. Firstly, divided the available network information in to two sets; one set is known as training set E^T, that contains the edges that is available as an input to the algorithm the remaining edge set is known probe set E^P (unknown information). That contains the edge set that is not available as an input. Secondly when algorithm gives set of predicted edges by taking training edge set as an input, the links are predicted by deciding a threshold value, the newly predicted generated edges are compared with probe set. That gives a ratio of correctly predicted edges. It is a similarity based algorithm, where each non-observed link between, x and y is assigned a score S_{xy} according to distance and number of overlapping neighbors. For some of the networks it can work wonderfully but for others it can be a failure. Moreover, as the algorithm is randomized, executing algorithm for each dataset only once is not the best way to obtain reliable results.

Pan et al [21], introduced an algorithmic structure where probability score of network's links is computed according to a pre-defined structural Hamiltonian. The non-observed link is scored and added to the network by considering a common standard called clustering coefficient, which states that a pair of node shows more chances of introducing link between them if some neighbors are common among them or are connected by short paths. This method makes use of high clustering coefficient of dissimilar networks.

Liao et al [22], proposed algorithm to compute the link score between pair of nodes based on similarity between nodes. The similarity is determined by the Pearson correlation coefficient. When applied to sum up the similarity which is based on high order paths, this method is seem a vigorous one. It’s been perceived that in simulation that score it-self cannot overtake the typical similarity measure such as CN, Jaccard, RA and Link Prediction methods.

Ibrahim and Chen [17], introduced ITM (Integrated time series model). A dynamic link prediction algorithm that is founded on the principle that newly introduced links can arose as well as present links can be wiped out among users of social network. Predicts future links in innovative social networks in four steps. Step 1 nodal information is converted in adjacency matrix with every time stamp. Step2 sequence of adjacency matrices $At_0, At_0 + 1, \ldots, At_0 + T - 1$ maintained over time is reduced to a weighted matrix, self-loops are also considered. If a link appeared in past then its occurrence probability can exploit the occurrence of two link pairs. The frequency of occurrence of link along with time is also considered. A frequently occurring link over a session may occur in upcoming time. Step3 communities are detected using improved reduced matrix weighted matrix generated in Step 2. Communities are discovered in two phases firstly modularity is optimized locally to detect small communities. Then nodes are combined which are associated with identical community in to a super node as well as construct network that may give a picture of communities. Then the super-nodes are merged within greater communities greater communities in order to have greater modularity temporal link information is incorporated with community information. Communication between individuals from different communities are also considered for future link prediction. Step 4 Node importance in a network is computed by considering greatest eigenvector results centrality. Topological information is merged with centrality and community structure. Step 5 the model of time series which uses temporal as well as topological information, three types of information, community information, node centrality information, and time series is combined. In this iterative algorithm, computation of initial local optimized state is very time consuming and have time complexity O^2. Same task can be accomplished by algorithms those runtime is $O(n.log2\,n)$.

Zhou et al [4], empirically investigated a simple back ground of link prediction for similarity measures on behalf of nine invoked similarity measures. Perceived from the experimentation results that the meekest similarity measure, which is named as Common Neighbors is seen to perform the best. A new similarity measure designed in which similarity scores are computed by making use of information on the next nearest neighbors,

For the set purpose a new measure is designed, in which the next nearest neighbor's information is used. It can stop the consistency of states and thus improve the algorithmic accuracy remarkably. The designed algorithm still needs further improvement.

Murata and Moriyasu [23] described an algorithm for predicting links in QABB (question answer bill board). That is based on both node attributes and structural properties of existing network. Recognizing the structure of previous communication related questions are suggested for integrated answers. Another example is to assume upcoming questions that will fascinate users.

Weights of links between users related to the number of times they encounter or impart on QABB. A social network is originated by putting links to all the pairs of the answerer's questions in each category. Famous QABB Yahoo! Chiebukuro (Japanese Yahoo! Answers) are used in testing.

Link weight score S_{xy} is allotted to each pair of nodes x and y and common neighbor algorithm is combined with the upper in-between node weights rather than the lower one. Encrypted user ID is used as an input e.g. categories, date and time. Only links among users who already exist in training period are the object for link prediction. Whole QABB data is divided within in groups and generates a social network for each category.

4. Experimental Setting

A. Methodology

For investigating the performance of algorithm we enquiry the stability of algorithm using 5 dissimilar datasets. We split the edge set E in to probe set E^P and training set E^T i.e., $E = E^T + E^P$. The ratio of link set between E^T and E^P is 80% and 20%. The probe set E^P is selected randomly from the total link set E, and is kept fixed for all algorithms. Algorithms are executed against E^P for each data set. Each under observation similarity based algorithm computes score if score equals or exceeds threshold value a link is formed between the two nodes. As a single execution might result in a biased result, therefore, each algorithm is executed 5 times on data sets, and threshold value for link formation between two nodes is computed by taking average of the available training set scores E^T of 80% location. Now to check the correctly predicted links percentage in each algorithm TP and FP is computed along with AUC to support the drawn results.

B. Algorithms

Common Neighbor and Distance

Common neighbor and distance algorithm is established on two structural properties of complex network, i.e., common neighbor and distance. The links are firstly scored against common neighbors between nodes $\frac{CN_{ij}+1}{2}$. CN_{ij} is the number of common neighbors between node i and j. When there are no common neighbors between them $\Gamma(i) \cap \Gamma(j) = \emptyset$, distance d_{ij} between nodes is used for score computation [20].

$$s_{ij} = \left\{ \begin{array}{ll} \frac{CN_{ij}+1}{2}, & \Gamma(i) \cap \Gamma(j) \neq \emptyset, \\ \frac{1}{d_{ij}}, & otherwise, \end{array} \right\}$$

Common Neighbor

In Common Neighbor Algorithm the score for link prediction is computed by finding the number of common neighbors between two distinct nodes, where $\Gamma(i)$ represents the neighbor of node i and $\Gamma(j)$ represents neighbors of node j [24].

$$s_{ij} = |\Gamma(i) \cap \Gamma(j)|$$

Preferential Attachment

In Preferential Attachment Algorithm the degrees of two distinct nodes are multiplied $k_i * k_j$ in order to compute score for link prediction. $k_i$is degree of node i and k_j is degree of node j [24],

$$s_{ij} = k_i * k_j$$

Sorensen

In Sorensen algorithm the twice of common nodes is divided on the product of degrees of two distinct nodes [24].

$$s_{ij} = \frac{2|\Gamma(i) \cap \Gamma(j)|}{k_i + k_j}$$

C. Data Sets

The real world complex network data sets that used for testing and competitive analysis.

The dataset are essential for comparison and reproduction of link prediction procedures. Gathering a valid dataset is time-consuming process and labor-intensive, as most of the datasets are not available publicly. We selected 5 popular datasets and used for link prediction, shown in Table 1 Particularly all the datasets represents real-world complex networks.

Future possible interaction between two parties is tremendously popular in recent times. In this paper link prediction algorithms are validated on the following real world complex network datasets.

1. USAir-The network of the US air transportation system, which contains 332 airports and 2126 airlines which connects the US around the globe [25].
2. Dolphins: It is network investigated by Lusseau et al. The network of 62 bottlenose dolphins who live in Doubtful Sound of New Zealand between 1994 and 2001 [26].
3. Karate: Data set of Zachary Karate club network, which shows the correlation of 34 members of a university Karate club. Dataset firstly studied by Wayne W. Zachary for over three years from 1970 to 1972 to study the clash arose between instructor and administrator [27].

(See data set from http://users.csc.calpoly.edu/~dekhtyar/466-Fall2010/labs/lab07.html).

4. Email: This is a network of e-mail exchanges between members of the Universitat Rovira i Virgili (Tarragona) [16].
5. Neurals: This data symbolizes the C. Elegans neural network of Graph is been processed in order to remove repeated edges [28]. (See data set from http://wormwiring.org/).

Network	N	m	c	<d>	<k>	cn
Karate	34	78	0.571	2.408	4.588	0.8590
Dolphins	62	159	0.259	3.357	5.129	0.7610
Pollbook	105	441	0.488	3.079	8.400	0.9592
E-mail	1133	5451	0.220	3.606	9.622	0.7758
Neurals	306	2147	0.292	Inf	14.0327	0.9446

Table 1 Illustration of properties of networks. N: number of nodes in G, m: number of edges in G, c: clustering coefficient, $< d >$: average distance, $< k >$: average degree, cn: CN coefficient

D. Evaluation Criterion

Three performance evaluation matrices are used to estimate the quality of the results after each algorithm calculates and scores similarities, A link prediction algorithm assign Score S_{xy} to every missing link (i.e., $U - E^T$) and provides an ordered list of edges according to their relevant score. A score S_{xy} is to quantify the existence likelihood of a missing link. If S_{xy} exceeds threshold value the link is validated and considered to be exist.

Three standard matrices are used to measure accuracy of prediction algorithm we evoke the TP (True Positive) checks correctly predicted links in E^P or validation set, FN (False Negative) the links that don't even predicted using E^T but are found in E^P. TP and FP are the common and robust accuracy validators. Area-under-curve (AUC) value understood as the probability that a randomly chosen existing link is given a higher score than a randomly chosen non-existent link. At each time, we randomly pick an existing link and a non-existent link in next time to match their scores. If among n independent comparisons, existing links have a higher score n1 times and the same score n2 times [29]. The value of AUC is computed by $AUC = (n1 + 0.5n2)/n$

We used AUC for accuracy checking because dissimilarity is much more prominent in terms of AUC. It's computed over all possible node pairs, not only node pairs without edges. Thus we support AUC to evaluate the accuracy of new link prediction. Remarkable property of AUC is its definition, it is the probability of positive example which is selected at random, and appears above a randomly selected negative sample [20].

5. Results And Discussions

We firstly compare the performance of four similarity based algorithms Common Neighbor and Distance, Common Neighbor, Preferential Attachment and Soreson Index on five representative data sets: Neurals, USAir, Email, Karate and Dolphin. Different algorithms gave diverse results on different datasets. The detailed AUC values of data sets are reported in Fig 3-7. Depicted values are average of 5 tests. We observe in our simulation that Soreson index performed very poorly. It cannot overtake the typical similarity method such as CN, CN and Distance, as well as Preferential Attachment in link prediction. Soreson Index shown paltry result for USAir network with an average TP of 68% and Standard deviation of 5. The algorithm produced unsatisfactory link prediction percentage for Neural network with TP of 56% and with standard deviation of 9. Preferential Attachment algorithm is placed on third in term of performance. Common Neighbor algorithm enjoys a high prediction accuracy and can precisely predict links in complex networks. It is scored second best algorithm for link prediction. CN and distance algorithm regarded as best link predictor among all link predictors, CN and distance algorithm shown adequate performance on neural and dolphin network. The results are also reported in Figure 3 and 7.

We regard common neighbor and distance algorithm as improved version of CN algorithm. That combines distance measure for prediction. We perceived 82% average TP in karate network with standard deviation of 1.9.

	CN		CN & Distance		Preferential Attachment		Soreson Index	
	Ave TP	Ave FP	Ave TP	Ave FP	Ave TP	Ave FP	Ave TP	Ave FP
Karate	68%	33%	83%	17%	45%	55%	68%	32%
Dolphins	70%	30%	83%	17%	58%	42%	70%	30%
E-mail	60%	40%	72%	28%	55%	45%	60%	40%
USAIR	75%	25%	68%	32%	58%	42%	75%	25%
Neurals	50%	50%	56%	44%	44%	56%	50%	50%

Table 2 TP and FP Percentage

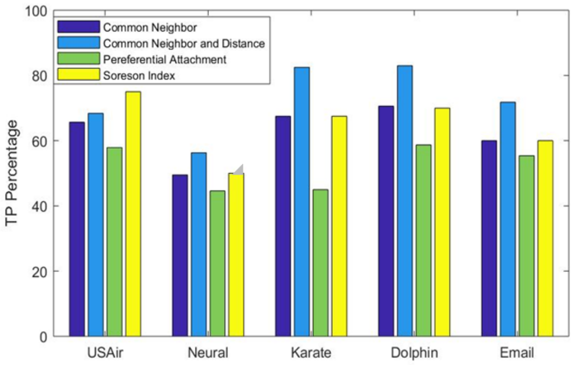

Figure 5 TP and FP

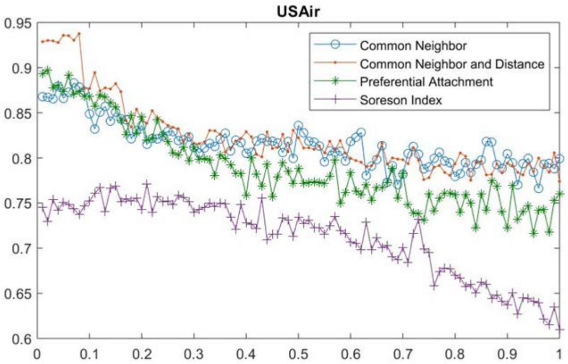

Figure 5 Comparative Analysis of USAir Dataset

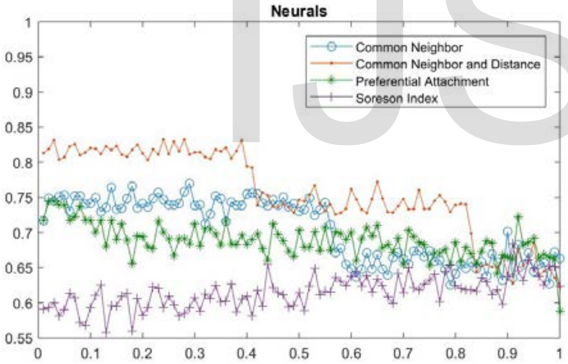

Figure 5 Comparative Analysis of Neurals Dataset

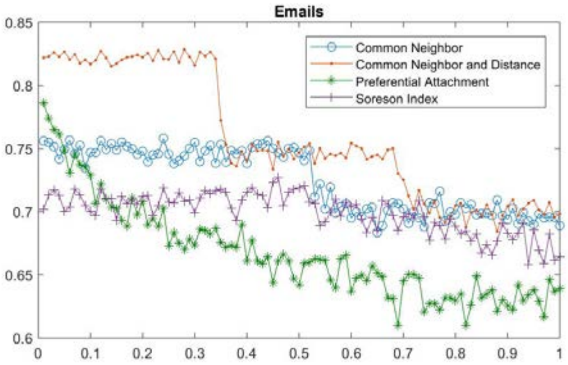

Figure 5 Comparative Analysis of Emails Dataset

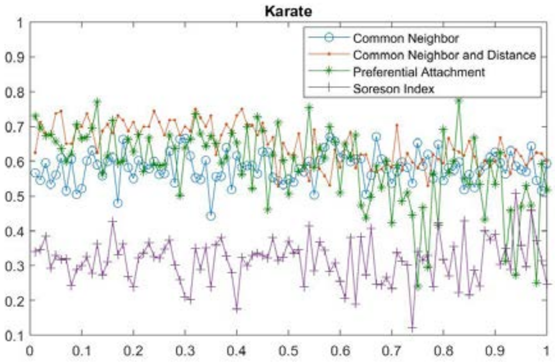

Figure 7 Comparative Analysis of Karate Dataset

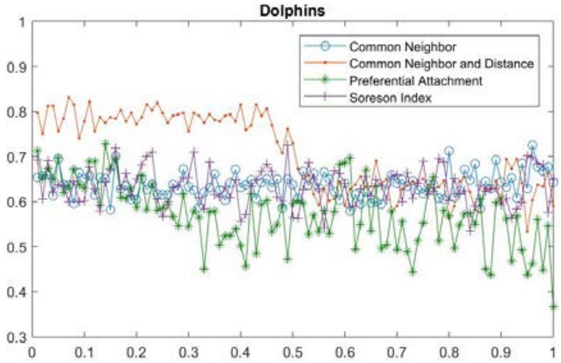

Figure 7 Comparative Analysis of Dolphins Dataset

Dolphins network gave an average of 83% TP with a standard deviation of 1.1, Email network resulted performance with an average of 71% TP in CN and Distance. Fig 2 shows the percentage of correctly predicted links True positive (TP) using 20% E^P respectively. Algorithm gave discrete results for each dataset. The TP percentage of four similarity based Algorithms Common Neighbor and Distance, Common Neighbor, Preferential Attachment and Soreson Index on five representative data sets and result shown in Table 2.

The crucial difficulty for the link prediction in sparse network is low degree nodes which can be tackled by incorporating more information. However, the AUC of CN and distance also decreases when sampling space is increased for AUC computation can be seen in Fig 3 and Fig 7.

6. Conclusion

In this paper four different link prediction algorithms are analyzed, based on similarity between nodes and concluded. We compared the four similarity based Algorithms Common Neighbor and Distance, Common Neighbor, Preferential Attachment and Soreson Index on five representative data sets and result shown that CN and distance algorithms outperforms in link prediction.

Many problems remain still open. The link prediction problem has not been completely solved. A detailed study of their performance in link prediction would be another addition. For example, how to choose a suitable link prediction algorithm according to properties of network as there is no absolute method for all networks. Our results showed that adding more information properly can increase the accuracy and percentage of missing link prediction.

References

1. Newman, M.E., *The structure and function of complex networks.* SIAM review, 2003. **45**(2): p. 167-256.
2. Wang, T., et al., *Link prediction in evolving networks based on popularity of nodes.* Scientific reports, 2017. **7**(1): p. 7147.
3. Leroy, V., B.B. Cambazoglu, and F. Bonchi. *Cold start link prediction*. in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010. ACM.
4. Zhou, T., L. Lü, and Y.-C. Zhang, *Predicting missing links via local information.* The European Physical Journal B, 2009. **71**(4): p. 623-630.
5. Al Hasan, M., et al. *Link prediction using supervised learning*. in *SDM06: workshop on link analysis, counter-terrorism and security*. 2006.
6. Yang, Y., et al. *Link prediction in human mobility networks*. in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. 2013. IEEE.
7. Kushin, M.J. and M. Yamamoto, *Did social media really matter? College students' use of online media and political decision making in the 2008 election.* Mass Communication and Society, 2010. **13**(5): p. 608-630.
8. Lü, L. and T. Zhou, *Link prediction in complex networks: A survey.* Physica A: statistical mechanics and its applications, 2011. **390**(6): p. 1150-1170.
9. Boccaletti, S., et al., *Complex networks: Structure and dynamics.* Physics reports, 2006. **424**(4-5): p. 175-308.
10. Dorogovtsev, S.N. and J.F. Mendes, *Evolution of networks.* Advances in physics, 2002. **51**(4): p. 1079-1187.
11. Getoor, L. and C.P. Diehl, *Link mining: a survey.* Acm Sigkdd Explorations Newsletter, 2005. **7**(2): p. 3-12.
12. Gong, N.Z., et al., *Joint link prediction and attribute inference using a social-attribute network.* ACM Transactions on Intelligent Systems and Technology (TIST), 2014. **5**(2): p. 27.
13. Gupta, P., et al. *Wtf: The who to follow service at twitter*. in *Proceedings of the 22nd international conference on World Wide Web*. 2013. ACM.
14. He, Y.-l., et al., *OWA operator based link prediction ensemble for social network.* Expert Systems with Applications, 2015. **42**(1): p. 21-50.
15. Redner, S., *Networks: teasing out the missing links.* Nature, 2008. **453**(7191): p. 47.
16. Guimera, R., et al., *Self-similar community structure in a network of human interactions.* Physical review E, 2003. **68**(6): p. 065103.
17. Ibrahim, N.M.A. and L. Chen, *Link prediction in dynamic social networks by integrating different types of information.* Applied Intelligence, 2015. **42**(4): p. 738-750.
18. Albert, R. and A.-L. Barabási, *Statistical mechanics of complex networks.* Reviews of modern physics, 2002. **74**(1): p. 47.
19. Schafer, J.B., J.A. Konstan, and J. Riedl, *E-commerce recommendation applications.* Data mining and knowledge discovery, 2001. **5**(1-2): p. 115-153.
20. Yang, J. and X.-D. Zhang, *Predicting missing links in complex networks based on common neighbors and distance.* Scientific reports, 2016. **6**: p. 38208.
21. Pan, L., et al., *Predicting missing links and identifying spurious links via likelihood analysis.* Scientific reports, 2016. **6**: p. 22955.
22. Liao, H., A. Zeng, and Y.-C. Zhang, *Predicting missing links via correlation between nodes.* Physica A: Statistical Mechanics and its Applications, 2015. **436**: p. 216-223.
23. Murata, T. and S. Moriyasu, *Link prediction based on structural properties of online social networks.* New Generation Computing, 2008. **26**(3): p. 245-257.
24. Lü, L., C.-H. Jin, and T. Zhou, *Similarity index based on local paths for link prediction of complex networks.* Physical Review E, 2009. **80**(4): p. 046122.
25. Sett, N., S.R. Singh, and S. Nandi, *Influence of edge weight on node proximity based link prediction methods: An empirical analysis.* Neurocomputing, 2016. **172**: p. 71-83.
26. Lusseau, D., et al., *The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations.* Behavioral Ecology and Sociobiology, 2003. **54**(4): p. 396-405.
27. Silva, T.C. and L. Zhao, *Semi-supervised learning guided by the modularity measure in complex networks.* Neurocomputing, 2012. **78**(1): p. 30-37.
28. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world'networks.* nature, 1998. **393**(6684): p. 440.
29. Jiang, M., Y. Chen, and L. Chen, *Link prediction in networks with nodes attributes by similarity propagation.* arXiv preprint arXiv:1502.04380, 2015.