Scientific Programming manuscript no. output
December 1, 2014

Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics

Fei Gao[1], Katarzyna Musial[1], Colin Cooper[1], and Sophia Tsoka[1]

King's College London, School of Natural and Mathematical Sciences, Department of Informatics, Strand Campus, London, WC2R 2LS, United Kingdom
e-mail: fei.1.gao@kcl.ac.uk; katarzyna.musial@kcl.ac.uk; colin.cooper@kcl.ac.uk; sophia.tsoka@kcl.ac.uk

ABSTRACT

Currently, we are experiencing a rapid growth of the number of social–based online systems. The availability of the vast amounts of data gathered in those systems brings new challenges that we face when trying to analyse it. One of the intensively researched topics is the *prediction of social connections between users*. Although a lot of effort has been made to develop new prediction approaches that could provide a better prediction accuracy in social networked structures extracted from large–scale data about people and their activities and interactions, the existing methods are not comprehensively analysed. Presented in this paper, research focuses on the link prediction problem in which in a systematic way, we investigate the correlation between network metrics and accuracy of different prediction methods. For this study we selected six time–stamped real world social networks and ten most widely used link prediction methods. The results of our experiments show that the performance of some methods have a strong correlation with certain network metrics. We managed to distinguish 'prediction friendly' networks, for which most of the prediction methods give good performance, as well as 'prediction unfriendly' networks, for which most of the methods result in high prediction error. The results of the study are a valuable input for development of a new prediction approach which may be for example based on combination of several existing methods. Correlation analysis between network metrics and prediction accuracy of different methods may form the basis of a metalearning system where based on network characteristics and prior knowledge will be able to recommend the right prediction method for a given network at hand.

Key words. Social Network, Link Prediction, Network Metrics, Correlation

1. Introduction

Network structures have been studied for many years. First research in this area can be traced back to 1736 when Euler defined and solved the Seven Bridges problem of Königsberg [5]. Since then, for a long time, networks have been mainly studied by mathematicians and this resulted in a very prominent research field known today as the graph theory. There was not much ground breaking development in the complex network research area until 1960s, when the Erdos-Renyi random graph model (ER–model) was introduced [15, 16]. This is the simplest model of complex network. Due to the fact that there was a lack of large real world data, most of the work had been done on theoretical analysis of phenomena existing in networked structures (e.g. phase transition).

Over the years data collection techniques have significantly improved our ability to store massive and heterogenous network data. During the time when ER–model was introduced, progress has also been made by sociologists in researching real world human relationships [51, 36]. A new wave of research was set off by Watts and Strogatz who published a paper about the small–world effect in 1998 [54] and introduction of the scale–free network model by Barabasi and Albert one year later [4].

As the accessibility of database systems and Internet is growing, more and more real world network datasets are available. The available information about people and their activities is much richer and more complex than ever before. The complex network concept is an abstract form of various real–world networks, e.g. biological networks such as protein-protein interaction networks, metabolic networks [22, 26], human networks and disease spread [47, 53, 9, 11], scientific collaboration networks [21, 41] and online social networks [1, 13, 17, 23].

Link prediction in complex network is one of the popular research topics. Most of the researchers focus on the link prediction problem [31] which is very valuable for solving real world problems. Generally, the prediction problem is mainly studied from two angles: (i) network structure and (ii) attributes of nodes and connections. Structure refers to the way in which nodes that compose the network are interconnected. It reflects the information about network topology. Majority of the progress in the area of structure based prediction has been made by mathematicians and physicists. Some of the well–known structure based prediction methods are Common Neighbour, Jaccard's Index, Adamic/Adar Index, Katz, etc. (for a review of the methods please see [34]).

The link prediction problem also has been studied from the angle of the network attribute information. The attribute information refers to description of the features of nodes. Such information is difficult to show directly in the network graph. It can be for example done by labelling nodes, e.g. 1 depicts node that represents woman and 2 means that node represents man. The majority of attribute-based prediction methods follow a machine learning approach, i.e. they use classification–based methods to make predictions. Widely used methods include Decision Tree, Support Vector Machine(SVM), Naïve Bayes, etc. [45, 33]. In [32, 55, 20], authors report that the performance of link prediction improves when machine learning approaches are used. However, this is done using additional network information that

is not always available. We would like to emphasize that in our work, we are interested in the methods that only require the basic network structure information and thus we do not include machine learning methods in our study.

However, although much effort has been made, there is still no prominent prediction method that could provide a satisfactory performance. Thus, there is still a huge research gap that needs to be addressed.

1.1. Research Motivation

In the realm of network prediction, many efforts have been done on exploring new prediction methods that could provide better performance. However, the methods presented in most of the studies only improve the prediction result significantly for the network used in the study. There is a lack of systematically research that would enable to reveal the reason why the methods are good predictors when it comes to some of the networks but very bad when other networks are considered.

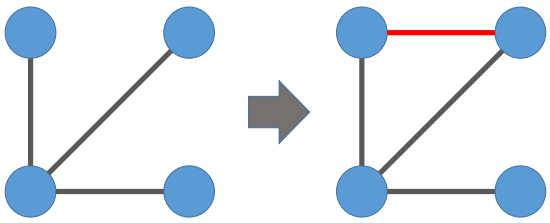
This paper addresses this problem, by exploring the correlation between network metrics and prediction accuracy of different methods. We expect that such approach will enable to find the reasons why methods performance vary on different networks. Apart from having a further understanding of the prediction methods, the study is also important as a theoretical base for developing new prediction methods. This could be relevant to many subjects. The prediction methods could help to find the relationships between proteins which might not be easily observed directly due to the interaction complexity . For example, new interactions can be inferred from the existing known interaction networks [10, 2] which shows a much better performance than prediction purely by chance. Online market targeting might also benefit from the network prediction which has already been applied in real world industries. For example, Google and Amazon recommend customers the potential goods and services that they might be interested in which is a kind of link prediction that predict the link between customers and products.

Beyond that, analysis of the link prediction problem in a time series approach could help researchers gain a better understanding of the evolution of the networks. Many works have been done to study the dynamic of complex network [6, 7, 8]. The achievement of network prediction analysis could help explain the mechanism of the network evolution.

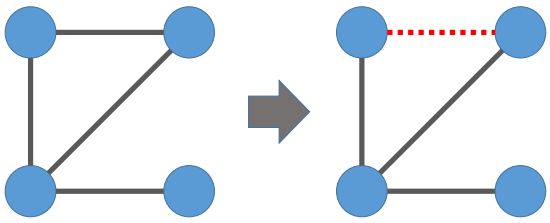
1.2. Contributions

The main contribution of our study is that we look at the link prediction as a time series problem and systematically analysed the correlation between network metrics and methods accuracy. In addition, in our experiments, we also find that for some networks, most of the prediction methods could provide a good performance while for some other networks, most methods are relatively powerless. We name them 'prediction friendly' networks and 'prediction unfriendly' networks respectively.

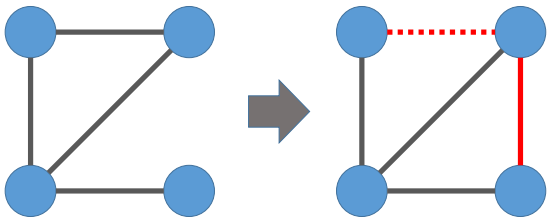
The paper is structured as follows: Section 2 presents the prediction methods and performance metrics used our experiments. Section 3 presents how the dataset were selected and processed. In section 4 and 5 we introduce the experimental design and present obtained results. We conclude the paper in Section 6.



(a) Adding Links



(b) Removing Links



(c) Adding & Removing Links

Fig. 1. Link Prediction Problems

2. Link Prediction Problem

Link prediction problem has been extensively studied by members of the complex network community. David Liben-Nowell and Jon Kleinberg have formalised the link prediction problem in [31] in the following way:

Let $G(V, L)$ be a network within the time period of $G[t, t_1]$ where V represents the set of nodes and L represents the set of links. For the next time period $G(t_1, t_2]$, the network might change. The link prediction focuses on how to predict the evolution of links, i.e. how $L_{[t,t_1]}$ will differ from $L_{(t_1,t_2]}$.

Researchers with background in physics and mathematics usually deal with the problem by focusing on the topology information of the networks. Researchers with machine learning and data mining background favour to solve the problem with considering the nodes' attribute information. There are three types of link prediction problems as shown in Fig 1: we can consider (i) only adding links to the existing network, (ii) only removing links from the existing structure, and (iii) both, adding and removing links at the same time.

Adding Links
Adding links (Fig 1a) means that in the next time window a new link will be created between existing nodes. There can be one or more newly created links.

Removing Links
Removing links (Fig 1b) means that the link will disappear in the next time window. Similar to the situation when new links are added, one or more more link can be removed in one time step.

Adding and Removing Links
This problem is the combination of two previously described problems. It means that from one time window to another both appearance and disappearance of links can be predicted (Fig 1c).

In this research, we will only focus on the first type of link prediction problem which only aims at predicting the appearance of links. The main reason for this is that the vast majority of existing methods for real–world data focus on this problem, so it means that we have big enough base to perform correlation analysis.

2.1. Prediction Methods

We select and present a brief description of ten commonly used prediction methods that use topology information about networks in the prediction process. Throughout this section the symbols x, y denote nodes, N denotes number of nodes in the network, and k is the average degree. $\Gamma(x)$ and $\Gamma(y)$ denote the neighbour sets of these nodes, k_x and k_y denote the degree number of node x and y respectively.

Common Neighbours
This method is based on the assumption that two nodes with many common neighbours will be connected in the future. The more common neighbours two users have, the higher the probability that a relationship between them will emerge. As a basic and intuitive method, Common Neighbours approach is usually used as a baseline to judge the performance of other methods [34, 31, 14, 13]. The complexity of this method, as introduced in [35], is $O(Nk^2)$.

$$|\Gamma(x) \cap \Gamma(y)| \tag{1}$$

Jaccard's Coefficient
The Jaccard's Coefficient, also known as Jaccard index or Jaccard similarity coefficient, is a statistic measure used for comparing similarity of sample sets. It is usually denoted as $J(x, y)$ where x and y represent two different nodes in a network. In link prediction, all the neighbours of a node are treated as a set and the prediction is done by computing and ranking the similarity of the neighbour set of each node pair. This method is based on Common Neighbours method and its complexity is also $O(Nk^2)$. The mathematical expression of this method is as follows [31]:

$$\left|\frac{\Gamma(x) \bigcap \Gamma(y)}{\Gamma(x) \bigcup \Gamma(y)}\right| \tag{2}$$

Preferential Attachment
Due to the assumption that the node with high degree is more likely to get new links [42], preferential attachment was introduced as a prediction method. The degree of both nodes in a pair needs to be considered for the prediction. Same as common neighbours, this is also a basic prediction method which is usually used as a baseline to measure the performance of other prediction methods. This method will calculate similarity score for each pair of nodes within the network rather than only the neighbour of nodes thus the complexity of Preferential Attachement is $O(N^2k^2)$. This method can be expressed as:

$$|\Gamma(x)| * |\Gamma(y)| \tag{3}$$

Adamic/Adar Index
It was initially designed to measure the relation between personal home pages. As shown in equation 4, the more friends z has, the lower score it will be assigned to. Thus, the common neighbour of a pair of nodes with few neighbours contributes more to the Adamic/Adar score (AA) value than this with large number of relationships. In real world social network, it can be interpreted as: if a common acquaintance of two people has more friends, then it is less likely that he will introduce the two people to each other than in the case when he has only few friends. It shows good results in predicting the friendship according to personal homepage and Wikipedia Collaboration Graph, but in the experiment of predicting author collaboration, it shows a poor accuracy prediction [1]. It is another method that is based on common neighbour, the complexity is also the $O(Nk^2)$. It is calculated as:

$$\sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{4}$$

Where z is a common neighbour of node x and node y.
Katz$_\beta$
This method takes lengths of all paths between each pair of nodes into consideration [24]. According to equation 5, the number of paths between node x and node y with length l (written as $|paths_{xy}^{\langle l \rangle}|$) are calculated and then multiplied by a factor β^l. By summing up all the results for a given two nodes with path length from 1 to ∞, a prediction score for the pair of nodes (x, y) is obtained. Katz is a prediction method based on the topology of whole network and thus its calculation is more complex than other methods in this section. The complexity is mainly determined by the matrix inversion operator, which is $O(N^3)$ [35, 18].

$$\sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{\langle l \rangle}| \tag{5}$$

The parameter β, as shown in equation 5, is used to adjust the weight of path with different length. When an extremely small β is chosen, the longer paths will contribute less to the score in comparison to shorter ones so that the result will be close to the common neighbours.

It is one of the prediction methods that, as it will be shown in further sections, achieves high prediction accuracy in many experiments.
Cosine Similarity
The idea of this method is based on the dot product of two vectors. It is often used to compare documents in text mining [34]. In network prediction problem, this method is expressed as:

$$\frac{|\Gamma(x)||\Gamma(y)|}{||\Gamma(x)|| * ||\Gamma(y)||} \tag{6}$$

For each pair of nodes with common neighbours, this methods will perform a vector multiplication and thus the complexity is $O(Nk^3)$.

Sørensen Index
This index [50] is designed for comparing the similarity of two samples and originally used to analysis plant sociology. The complexity of this method is $O(Nk^2)$. It is defined as:

$$\frac{2|\Gamma(x) \bigcap \Gamma(y)|}{k_x + k_y} \tag{7}$$

Hub Promoted Index
HPI is proposed for analysing metabolic networks as shown in [48]. The property of this index is that the links adjacent to hubs are likely to obtain a higher similarity score. The complexity of the method is $O(Nk^2)$. It is expressed as:

$$\frac{|\Gamma(x) \bigcap \Gamma(y)|}{min\{k_x, k_y\}} \tag{8}$$

Hub Depressed Index
Approach that uses the idea of hub in totally different manner than HPI is Hub Depressed Index (HDI). It gives links adjacent to hub a lower score. Its complexity is same as Hub Promoted Index, $O(Nk^2)$. It is defined as

$$\frac{|\Gamma(x) \bigcap \Gamma(y)|}{max\{k_x, k_y\}} \tag{9}$$

Leicht-Holme-Newman Index
LHNI [30] was proposed to quantify the similarity of nodes in networks. It is based on the concept that two nodes are similar if their immediate neighbours in the network are themselves similar. As another common neighbour based method, its complexity is $O(Nk^2)$.It is defined as:

$$\frac{|\Gamma(x) \bigcap \Gamma(y)|}{k_x * k_y} \tag{10}$$

All of the methods presented in this section are following similar approach. The required input for each method is the adjacency matrix that represents a network in which there are only 0 and 1 (0 – when there is no link between two given nodes, 1 – when the links between two given nodes exists). The output of each method is a similarity matrix in which each element represents the similarity score of a pair of nodes within the network and it is calculated according to the equation used in a given method.

2.2. Prediction Performance Metrics

In order to measure the performance of a prediction method, we need to use historical network data. Link prediction is a time related activity, therefore, we should use time–stamped dataset and according to the time stamp, separate the data into two sets, $G_{t,t_1}(V, L_1)$ as training set for prediction methods and $G_{t_1,t_2}(V, L_2)$ as unknown future network for testing where $t < t_1 < t_2$. Those two networks must consist of the same set of nodes V. The number of possible links that is denoted by U is $|V| * (|V| - 1)/2$. The link prediction method, in principle, provides a similarity score for each non existing links $(U - L_1)$ and for most methods, a higher score means higher likelihood that the link will appear in the future. Final prediction is done by ordering this score list and selecting top N links with the highest score.

In our work, AUC is used for quantifying the accuracy of prediction method. It is the area under the receiver operating characteristic curve [19]. In the context of network link prediction, AUC can be interpreted as the probability that a randomly chosen missing links $(L_1 \cup L_2 - L_1)$ is given a higher similarity score than a randomly chosen pair of unconnected links $(U - (L_1 \cup L_2))$ [12]. The algorithmic implementation of AUC follows the approach in [34]. It is calculated as

$$\frac{n' + 0.5n''}{n} \tag{11}$$

Where n is the number of times that we randomly pick a pair of links from missing links set and unconnected links set; n' is the number of times that the missing link got a higher score than unconnected link while n'' is the number of times when they are equal. The AUC value will be 0.5 if the score are generated from and independent and identical distribution. Thus, the degree to which the AUC exceeds 0.5 indicates how much better the predictions when compared to predict by chance.

3. Data Preparation

All six datasets used in experiments are real world social networks, five of them come from Koblenz Network Collection (KONECT [28]) and another from the Wrocław University of Technology. (see Table 1)

3.1. Dataset Selection

Datasets for the experiments have to meet certain requirements: (i) they have to represent data about users' interactions or any other type of activity that enables to define connections between users, and (ii) those activities have to be time stamped. As described in section 2, the link prediction problem is a time series problem that looks into the evolution of networks in time. Time-stamp is thus necessary. Table 1 shows the original dataset information that were selected based on these two criteria.

3.2. Data Processing

To make the data suitable for the experiments, first the preprocessing of datasets has been performed. It consists of the following three steps:

1. **Select data samples**. For each dataset, we first randomly select 6000 - 8000 user records (8000 samples is selected due to the calculation capacity. As for some dense networks, 8000 nodes is also too big, so we choose 6000) from the original dataset as the sample user data. As UC Irvine Messages only contains 1899 users, so we leave it as it is. The specific sample numbers are shown in Table 2.

2. **Split the data into training and testing sets**. Prediction in a time series problem means the dataset should be divided into train and test sets based on time stamps available. As the dataset of Flickr and YouTube are collected by taking snapshot of the network which is different from other four datasets, we take the first day snapshot as the training set and the remaining data as the test set. The other four networks are split according to the time scale with a ratio approximate training time : test time = 80% : 20% as shown in Table 2.

3. **Extract connected network**. Dividing data into training and testing sets can cause the isolation of some nodes or cliques. This, in turn, generates noise for measuring the accuracy of prediction methods as the methods we selected can not predict unconnected nodes. To eliminate the impact

Table 1. Original Dataset Information

Dataset Name	Time Range	Vertices	Edges
Enron E-mail Communication[a]	1998/11 - 2002/07	87,273	1,148,072
Facebook Wall Posts[b]	2008/01 - 2009/01	63,731	1,269,502
Flickr Friendship[c]	2006/11 - 2007/05	2,302,925	33,140,018
PWr E-mail Communication[d]	2008/11 - 2009/05	14,316	49,950
UC Irvine Messages[e]	2004/03 - 2004/10	1,899	59,835
YouTube Friendship[f]	2006/12 - 2007/07	3,223,589	12,223,774

Notes. This table shows the original information about the datasets used in the experiments.
[(a)] The Email network among employees of Enron. Nodes in the network are individual employees and edges are individual emails [27].
[(b)] The wall posts from the Facebook New Orleans networks [52].
[(c)] The social network of Flickr users and their friendship connections. It is collected by taking a snapshot of the network on November 2nd, 2006 and record it daily until December 3rd, 2006, and then again daily between February 3rd, 2007 and May 18th, 2007[39, 40].
[(d)] The Email communication of Wrocław Univerisity of Technology [25].
[(e)] The network contains messages send between the users of an online community of students from the University of California, Irvine. A node represents a user. An edge represents sent message. Multiple edges denote multiple messages [46].
[(f)] The social network of YouTube users and their friendship connections between December 10th, 2006 and January 15th, 2007, and again daily between February 8th, 2007 and July 23rd, 2007[37, 38].

Table 2. Dataset Details

Dataset Name	Train Time Range	Test Time Range	Sample Nodes	Final Nodes
Enron E-mail Communication	1998/11 - 2001/12	2002/01 - 2002/07	8000	5208
Facebook Wall Posts	2008/01 - 2008/11	2008/12 - 2009/01	8000	5784
Flickr Friendship	Snapshot on 2006/11/02	2006/11/03 - 2006/12/03& 2007/02/03 - 2007/05/18	6000	5949
PWr E-mail Communication	2008/11 - 2009/04	2009/04 - 2009/05	8000	5208
UC Irvine Messages	2004/03 - 2004/08	2004/08 - 2004/10	1899	1666
YouTube Friendship	Snapshot on 2006/12/10	2006/12/11 - 2007/01/15& 2007/02/08 - 2007/07/23	6000	6000

Notes. The time range of train and test set, the number of sample nodes selected from the original dataset and number of nodes in the giant component which are used as the final nodes set for the experiment are presented in table.

of this noise, we extract the giant component from training dataset as our final training set $G_{t,t_1}(V, L_1)$. The final test set $G_{t_1,t_2}(V, L_2)$ is obtained by extracting the network with all the nodes that exist in $G_{t,t_1}(V, L_1)$ from the original test set obtained from step 2. For nodes existing in the final training set but not present in the original test set, we just keep and leave them isolated in the final test set as it is formed by link disappearing.

After all, we get the train set $G_{t,t_1}(V, L_1)$ and test set $G_{t_1,t_2}(V, L_2)$ as described in section 2.2 where the both sets have same nodes V.

4. Experimental Design

In order to be able to apply all selected methods and taking into account the types of datasets available, the network is represented as a binary un-weighted network. This enables consistent and comprehensive review of the existing methods.

First, the prediction methods described in section 2.1 will be applied to each of the processed training sets to get the similarity matrix as the prediction result. The prediction results will be then evaluated using the testing set and the AUC for each method will be calculated.

For the implementation of those methods, we applied the toolbox that presented in [34] and all the experiments were implemented in Matlab.

As stated before, the main goal of the research is to explore the correlations between the accuracy of different prediction methods and network metrics. For the training set of each network, the network metrics are calculated with toolboxes provided by KONECT [28] and MIT Strategic Engineering research group. The metrics we calculate include:

Global Clustering Coefficient
It is defined in [43] as:

$$GCC = \frac{3 * \text{number of triangles in the network}}{\text{Number of connected triples of vertices}} \quad (12)$$

It shows the transitivity of the network as a whole. The coefficient range is between 0 and 1.

Average Clustering Coefficient [54]
It is based on local clustering C_l. For each of the vertex l, its local clustering coefficient can be calculated by:

$$C_l = \frac{\text{Number of triples connected to vertex } l}{\text{Number of triples centered on vertex } l} \tag{13}$$

and then the ACC can be calculated as:

$$ACC = \frac{1}{v}\sum_{l} C_l \tag{14}$$

where v is the number of nodes in a network.

Network Density
The ratio between existing links and all possible links given the node numbers.

$$\text{Network Density} = \frac{\text{Number of Existing Links}}{\text{Number of all possible links}} \tag{15}$$

where

$$\text{Number of all possible links} = \frac{v * (v-1)}{2} \tag{16}$$

Where v is the number of nodes in the network.

Gini Coefficient [29]
In network study is defined as:

$$G = \frac{2\sum_{i-1}^{n} id_i}{n\sum_{i-1}^{n} d_i} - \frac{n+1}{n} \tag{17}$$

where $d_1 \leq d_2 \leq d_3 \leq \cdots \leq d_n$ is the sorted list of degrees in the network and n is the number of nodes in a network. Its value is between 0 and 1, where 0 denotes total equality between degrees and 1 denotes dominance of single node.

Diameter
The longest of the shortest paths in the network.

$$Diameter = max_{i,j} d(i,j) \tag{18}$$

Where $d(i, j)$ is the shortest path between node i and j.

Average Shortest Path
The average number of the shortest paths between each pair of vertices.

$$ASP = \frac{1}{v \cdot (v-1)} \cdot \sum_{i!=j} d(i,j) \tag{19}$$

Once the accuracy of prediction for each method and the metrics for each network are calculated, the correlation between them will be analysed. The Pearson's Coefficient [49] is used to measure the correlation between accuracy of network prediction method and selected network metrics. It is a widely used statistic method to measure linear correlation between two variables, say W and Z. It is calculated as:

$$\frac{\sum_{i=1}^{n}(W-\bar{W})(Z-\bar{Z})}{\sqrt{\sum_{i=1}^{n}(W_i-\bar{W})^2}\sqrt{\sum_{i=1}^{n}(Z-\bar{Z})^2}} \tag{20}$$

The coefficient value is between -1 and 1 where -1 means that two variables are negatively linearly correlated and 1 means that they are positively linearly correlated.

Table 3. Theoretical GCC & ASP of Random, Real and Regular Network

	Random Network	YouTube	Regular Network
Nodes	6,000	6,000	6,000
Links	54,596	54,596	54,596
GCC	0.0030	0.0286	0.7064
ASP	2.9983	3.0709	164.8500
		UC Irvine	
Nodes	1,666	1,666	1,666
Links	11,582	11,582	11,582
GCC	0.00835	0.0197	0.6919
ASP	2.8186	3.0463	59.9108
		PWr	
Nodes	6,335	6,335	6,335
Links	15,334	15,334	15,334
GCC	0.0008	0.0048	0.5547
ASP	5.5499	4.0162	654.3060
		Flickr	
Nodes	5,949	5,949	5,949
Links	387,719	387,719	387,719
GCC	0.0219	0.0658	0.7442
ASP	1.7845	2.3447	22.8198
		Facebook	
Nodes	5,784	5,784	5,784
Links	14,507	14,507	14,507
GCC	0.0009	0.0341	0.5633
ASP	5.3717	5.7235	576.5205
		Enron	
Nodes	5208	5208	5208
Links	23977	23977	23977
GCC	0.0018	0.0290	0.6586
ASP	3.8548	3.6818	282.8037

5. Experiment Result

5.1. Network Profiles

The values of network metrics for each of the extracted social network are presented in Table 5. As it is much easier to set up relationship between people in online social network than in real world network, the average shortest path in our experiments are all smaller than six, the number suggested by the six degrees of separation theory [44]. The average shortest path of the six

Table 4. Analytical formulas for GCC & ASP in random and regular networks

	Random Network	Regular Network
GCC	$\frac{k}{v}$	$\frac{3(k-2)}{4(k-1)}$
ASP	$\frac{\log v}{\log k}$	$\frac{v}{2k}$

Notes. k is the average degree and v is the number of nodes in the network

selected networks is 3.65. This reflects the small–world property of the networks. People are closer to each other in online social networks than in face–to–face networks. This phenomenon was also pointed out in [3] where authors established that the average shortest path of Twitter is 3.43.

The degree distributions of the six networks, shown in Fig 3, indicates that they are scale-free networks as the distributions follow the power law.

We also compared the GCC and ASP metrics of the real network with the theoretical metrics of random network and regular network that have same number of nodes and links. The analytical formulas for GCC and ASP in random and regular networks with a given number of nodes and links are given in Table 4. The results of calculations for each analysed network are presented in Table 3.

Fig 2 plots the metrics of six analysed networks and related theoretical networks respectively. It shows that the clustering coefficient of the analysed networks are all between random and regular networks. Meanwhile, the average shortest path of real-world networks are all very close to the random networks. This two phenomena indicate the small–world property of analysed structures. Taking into account both metrics and node degree distribution, it can be concluded that those networks are a combination of small–world and scale–free networks.

5.2. Prediction Results

The prediction results are summarised in Table 6. Katz method achieved the best average performance and the overall performance is ranked as: Katz > Preferential Attachment > Adamic-Adar > Common Neighbours > Cosine Similarity > Jaccard Index > Hub Depressed Index > Hub Promoted Index > Sørensen > Leicht–Holme–Newman Index. By comparing the variance of each method, we find that the Katz also provides the most stable prediction performance among those methods while Common Neighbours is the worst performing approach. Overall, we find that Katz and Preferential Attachment provide good prediction accuracy together with a relatively stability.

To study the prediction results from the perspective of each network please see Figure 4. The prediction results of different methods align on the vertical lines for each network respectively. From this figure, we find that for some networks, most of the prediction methods could provide a good prediction result. Such networks include Flickr, Enron and YouTube. We call this type of networks the 'prediction friendly' network. Apart from this type of network, there are also some networks for which most of the prediction approaches provide fairly low accuracy, such as Facebook, UC Irvine and PWr. Similarly, we call those network 'prediction unfriendly' networks. Please note that in the experiments, for both prediction friendly and unfriendly networks, Katz$_\beta$ always provide a good performance level.

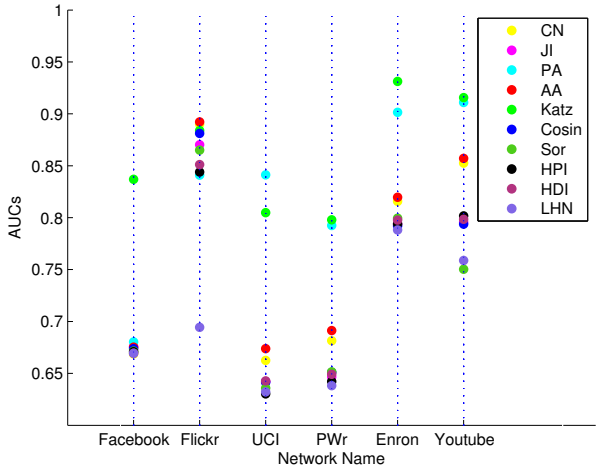


Fig. 4. The AUC Prediction Results for Each Network

5.3. Correlation between Prediction Accuracy and Network Metrics

Table 7 shows the Pearson's linear correlation coefficient of prediction accuracy and network metrics. The closer the absolute value to 1, the higher the correlation between analysed factors is. Figure 5 presents a heat–map plot to show the degree of linear relation between the two factors where we use the absolute value of Pearson's Coefficient. The brighter the colour in the heat–map is, the stronger a given network metric and the accuracy of prediction method are correlated.

In Figure 5, we can see that the Preferential Attachment and Gini Coefficient provides the highest correlation coefficient (0.94) which indicates that they generally follow a linear relationship. This is not a surprise. For a network with a high Gini Coefficient, there exist some nodes with dominant high degrees. It just reflects the phenomenon of "rich get richer" which is also the assumption of Preferential Attachment method. So we can say that preferential attachment could lead to a high Gini Coefficient and thus Preferential Attachment, on the other hand, could also describe how a network with high Gini Coefficient evolves by giving a better prediction result.

Cosine–GCC and Sor–GCC also provide a correlation coefficient above 0.8. We can draw the conclusion that Cosine Similarity and Sorensen Index method perform better in a network with higher GCC than it does in smaller GCC.

The Diameter and Average Shortest Path shows a negative linear relation to almost all of the prediction methods (excluding Katz and LHN where the negative correlation is weak). Both the Average Shortest Path and the Network Diameter reflect how easy it is to get from one node in a network to another one. Shorter path as well as smaller diameter means a higher probability that a pair of randomly picked nodes will be connected. Negative correlation between those two metrics and prediction accuracies of different methods means that most of the methods work well in the situations where networks feature short ASP and in consequence small Diameter. This is additionally supported by the fact that Global Clustering Coefficient is positively correlated with those of the prediction methods meaning that these methods work well with networks with high clustering

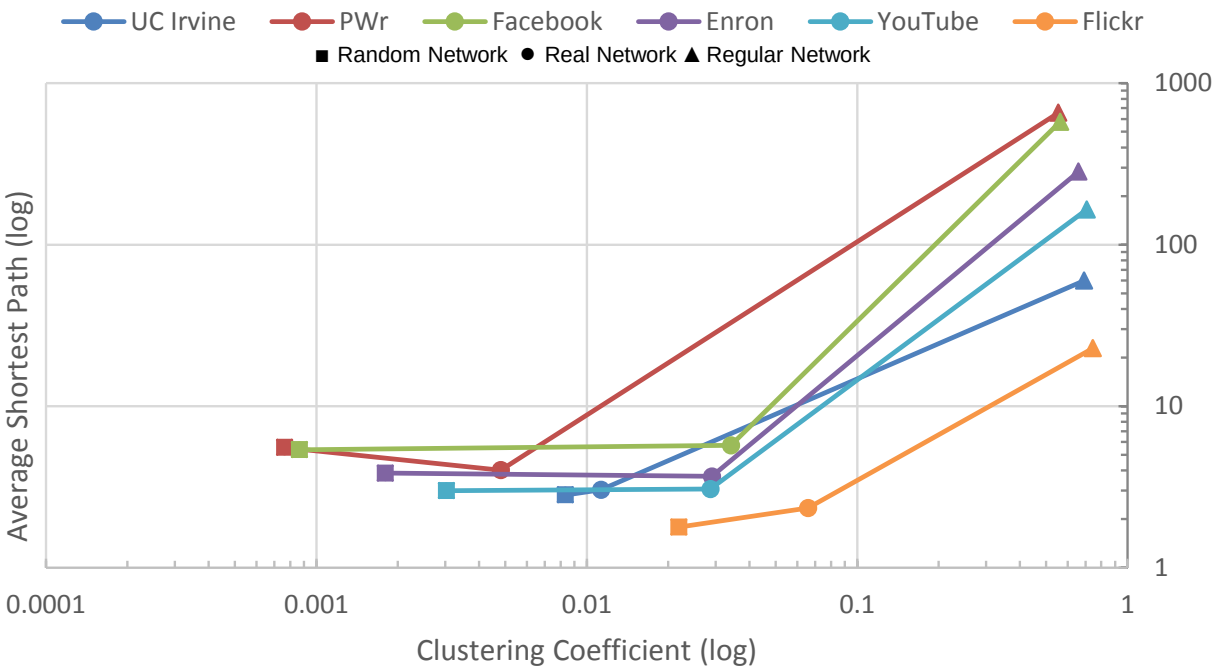


Fig. 2. Real Network and Theoretical Network Metrics Comparison

Table 5. Network Metrics Results

Datasets	GCC	ACC	Network Density	Gini Coefficient	Diameter	Ave Shortest Path
Facebook	0.0341	0.1176	0.0008674	0.473	16	5.7235
Flickr	0.0658	0.3294	0.0219	0.5931	6	2.3447
UC Irvine	0.0197	0.1075	0.0084	0.6394	7	3.0463
PWr	0.0048	0.2666	0.00076	0.6407	16	4.0162
Enron	0.029	0.1946	0.0018	0.7172	10	3.6818
YouTube	0.0286	0.2838	0.003	0.7222	5	3.0709

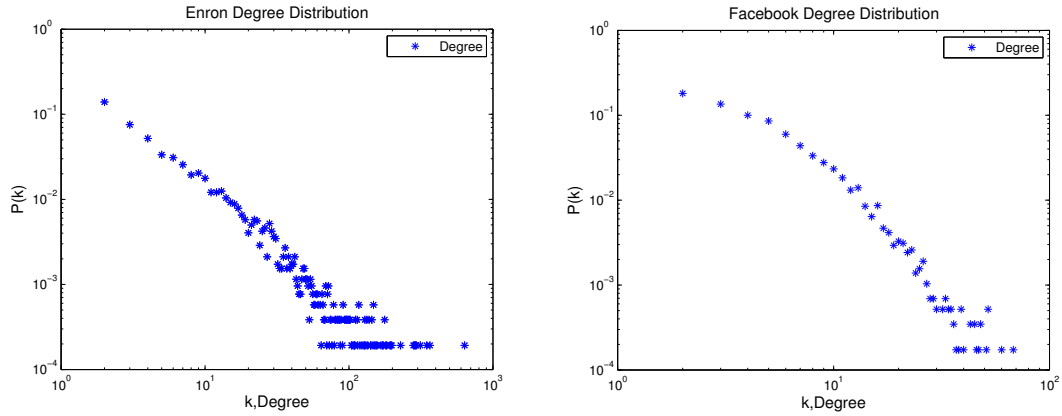
coefficient. Based on the above we can say that prediction methods positively correlated with GCC and negatively with ASP and Daimeter will work well in the situation where analysed network is of small–world type. In the same time they will work neither in random networks where GCC is very low nor in regular networks where ASP is very long.

It should be clear that the Pearson's coefficient does not indicate the accuracy of the method. For example, although the prediction method Katz does not show strong correlation to any of the network metrics, it still provides best result in our experiments. The reason can be found in Table 6, where it is shown that Katz always provides a high prediction accuracy regardless the tested network metrics.

The most important value of our correlation study lies in the variety of prediction methods used in the experiments. The prediction with methods combination could be a way to improve accuracy and this will be investigated in the future. The correlation between methods and network metrics could be used to determine the weight of different prediction methods in the combination process.

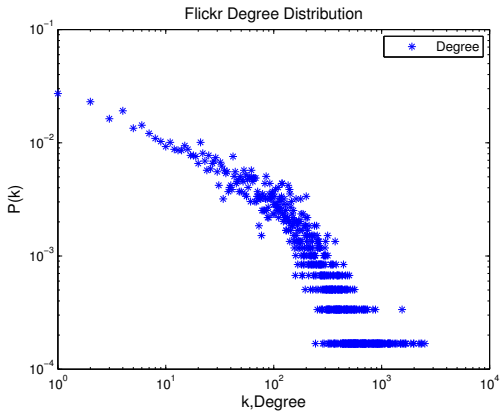
5.4. Prediction Friendly and Unfriendly Networks

Table 7 also shows the average correlation of network metrics and prediction accuracy. As we know the closer the absolute value of correlation to 1, the stronger the linear relation. Here we take 0.5 as a threshold for strong correlation. According to this, we find that there are four metrics strongly correlated with the prediction accuracy which includes GCC, ACC, Diameter and ASP. So it is reasonable to assume that these metrics could be used to classify the prediction friendly and unfriendly networks. We ranked each of the analysed networks according to the metrics that have strong correlation with prediction accuracy and based on this for each network we calculate the average ranking (Table 8). Top three ranked networks (with the small average ranks) are the prediction friendly networks and the other three are prediction unfriendly networks. It can be seen that the prediction friendly networks usually have large global and local clustering coefficient, a short average shortest path as well as small diameter. It suggests that networks with the structural profile similar to small–world network are easier to predict than networks similar to random structures.

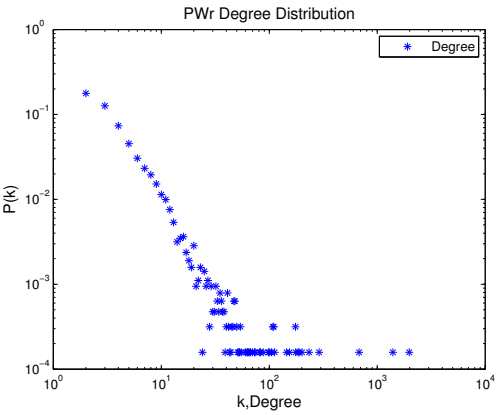


(a) Enron[a]

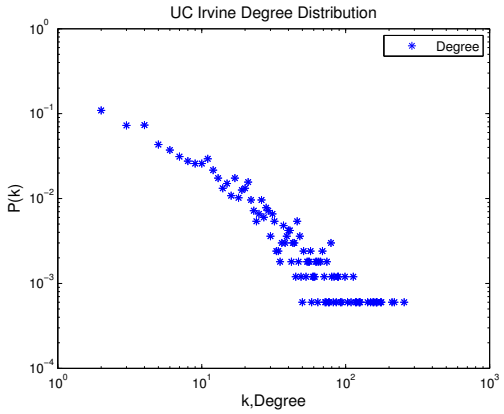
(b) Facebook[b]



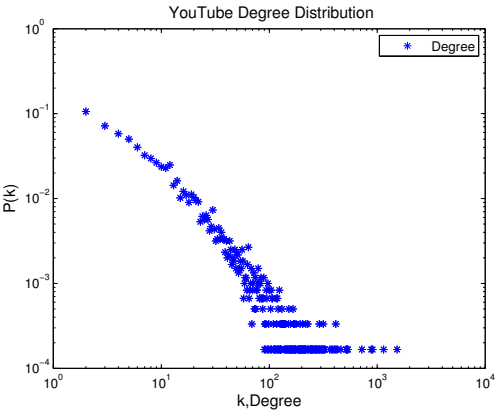
(c) Flickr[c]



(d) PWr[d]



(e) UC Irvine[e]



(f) YouTube[f]

Fig. 3. The Degree Distributions

Notes. The degree distributions are all follow the power law with exponent of :
[(a)] Enron, $r = 1.85$; [(b)] Facebook, $r = 1.82$; [(c)] Flickr, $r = 1.25$; [(d)] PWr, $r = 2.19$; [(e)] UC Irvine, $r = 1.56$; [(f)] YouTube, $r = 1.56$;

6. Conclusions

In this research, we look into the correlation between ten prediction methods and different network metrics in six time–stamped social networks. The study of network metrics confirmed that the node degree distribution of real world social networks follows a power law distribution. We also found that the average shortest path of online social network is much smaller than six. This might be due to the fact that online relationships are much easier to setup. The results of the prediction accuracy show that the best method among the tested ones is Katz$_\beta$. It is also the most stable technique from all tested ones. Preferential Attachment is the second best method that also provides a good prediction accuracy. In addition, for some 'prediction friendly' networks, most of prediction methods could provide a good performance while for some others, called in here as 'prediction unfriendly' networks, most prediction methods are lack of power.

Table 6. Prediction Methods Accuracy Result (AUC)

Datasets	AUC									
	CN	JI	PA	AA	$Katz_{\beta}$[a]	Cosin	Sor	HPI	HDI	LHN
Facebook	0.6688	0.6758	0.6803	0.6753	0.8369	0.6738	0.6715	0.6708	0.6694	0.6694
Flickr	0.89	0.8702	0.841	0.8922	0.8839	0.8812	0.865	0.844	0.8511	0.6944
UC Irvine	0.6625	0.6421	0.8412	0.6738	0.8048	0.6414	0.6359	0.6303	0.6427	0.6322
PWr	0.6815	0.6466	0.7924	0.6913	0.7979	0.651	0.6514	0.6422	0.6491	0.6382
Enron	0.8157	0.7937	0.9015	0.8196	0.9312	0.7921	0.7995	0.794	0.7977	0.7881
YouTube	0.8525	0.7957	0.9109	0.8571	0.9157	0.7938	0.7503	0.8017	0.7984	0.7587
Average	0.7618	0.7374	0.8279	0.7682	0.8617	0.7389	0.7289	0.7305	0.7374	0.6968
Variance	0.0105	0.0091	0.0071	0.0099	0.0032	0.0095	0.0084	0.0087	0.0083	0.0041

Notes. The accuracy of selected prediction methods measured by AUC. The average performance and the variance for each methods are also listed.
[(a)] In our experiment, we choose $\beta = 0.0005$

Table 7. Pearson Correlation of Prediction Methods Accuracy and Network Metrics

	CN	JI	PA	AA	$Katz_{\beta}$	Cosine	Sor	HPI	HDI	LHN	AVERAGE
GCC	0.68	0.79	0.05	0.68	0.47	0.80	0.81	0.73	0.74	0.27	0.60
ACC	0.75	0.68	0.43	0.76	0.39	0.70	0.65	0.67	0.68	0.30	0.60
Network Density	0.52	0.58	0.18	0.52	0.09	0.61	0.61	0.48	0.52	-0.12	0.40
Gini	0.45	0.30	0.94	0.46	0.49	0.29	0.25	0.36	0.37	0.57	0.45
Diameter	-0.67	-0.61	-0.77	-0.68	-0.51	-0.61	-0.52	-0.61	-0.63	-0.39	-0.60
ASP	-0.63	-0.55	-0.79	-0.65	-0.29	-0.57	-0.52	-0.52	-0.56	-0.18	-0.53

Notes. This table shows the correlation between prediction methods accuracy and network metrics calculated with Pearson's linear correlation coefficient. The number within the range of [-1,1] where 1 is completely positive correlation, 0 is no correlation, and -1 is completely negative correlation.

Table 8. Metrics Rank of Networks

Dataset	GCC	ACC	Diameter	ASP	Ave Rank
PWr	6	3	5	5	4.75
Facebook	2	5	5	6	4.5
UC Irvine	5	6	3	2	4
Enron	3	4	4	4	3.75
YouTube	4	2	1	3	2.5
Flickr	1	1	2	1	1.25

The Pearson correlation coefficient enabled to investigate the relationship between network metrics and prediction accuracy. Our research showed that some methods are highly correlated with certain network metrics (e.g. PA–Gini, Sor–GCC and Cosine–Gcc).

There are several further directions of the presented study. As discovered, for some networks, most prediction methods could provide a good performance which we name them as 'prediction friendly networks'. Similarly, we also find the existence of 'prediction unfriendly' networks. Section 5.4 explores the prediction friendly and unfriendly network classification according to the metrics ranking. The problem is that it does not provide an exact threshold that could be used to classify networks. It is out of scope of this research but is a very interesting topic for another study that we plan to conduct.

Based on the results of correlation between network metrics and the prediction accuracy, another possible work is to develop a new prediction approach which combine several, existing methods. We can also extend this research to many other networks, not only social ones, which might be good for finding some more general relations.

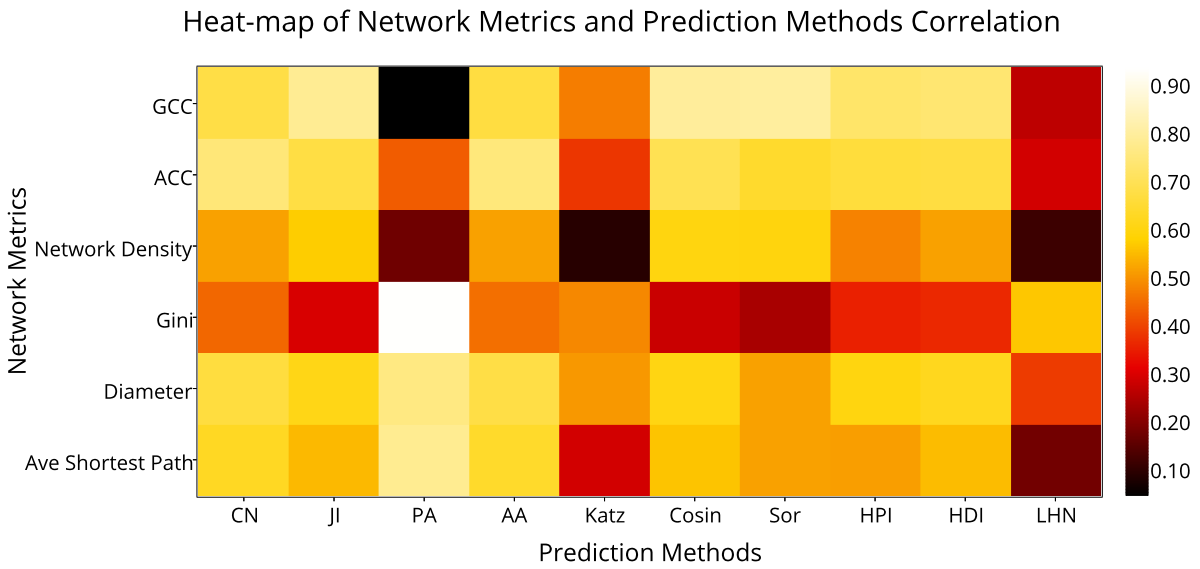


Fig. 5. Heat-map of Network Metrics and Prediction Methods Correlation

Notes. As for the Pearson Coefficient, both 1 and -1 stands for linear relationship (positive and negative), we use the absolute value of correlation coefficient in this figure to indicate whether the two factors are linearly correlated.

References

[1] Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

[2] Patrick Aloy and Robert B. Russell. Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162, 2003.

[3] Reza Bakhshandeh, Mehdi Samadi, Zohreh Azimifar, and Jonathan Schaeffer. Degrees of separation in social networks. In *SOCS*, 2011.

[4] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] N. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory, 1736-1936*. Clarendon Press, New York, NY, USA, 1986.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5):175 – 308, 2006.

[7] M. Budka, K. Juszczyszyn, K. Musial, and A. Musial. Molecular model of dynamic social network based on e-mail communication. *Social Network Analysis and Mining*, 2013.

[8] Guido Caldarelli, Alessandro Chessa, Irene Crimaldi, and Fabio Pammolli. The Evolution of Complex Networks: A New Framework. March 2012.

[9] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1:1–1:26, January 2008.

[10] Xue-Wen Chen and Mei Liu. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400, 2005.

[11] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[12] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[13] W. Cukierski, B. Hamner, and Bo Yang. Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244, 31 2011-Aug. 5.

[14] H.R. de Sa and R.B.C. Prudencio. Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2281–2288, 31 2011-Aug. 5.

[15] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[16] P. Erdős and A Rényi. On the evolution of random graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960.

[17] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80, Oct.

[18] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[19] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982.

[20] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[21] Zan Huang, Xin Li, and Hsinchun Chen. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 141–142, New York, NY, USA, 2005. ACM.

[22] Mason S. P. Barabasi A.-L. Oltvai Z. N. Jeong, H. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[23] K. Juszczyszyn, K. Musial, and M. Budka. Link prediction based on subgraph evolution in dynamic social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 27–34, 2011.

[24] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[25] Przemyslaw Kazienko, Katarzyna Musial, and Aleksander Zgrzywa. Evaluation of node position based on email communication. *Control and Cybernetics*, 38(1):67–86, 2009.

[26] A. D. King, N. Pržulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, November 2004.

[27] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proc. European Conf. on Machine Learning*, pages 217–226, 2004.

[28] Jérôme Kunegis. Konect: the koblenz network collection. In *WWW (Companion Volume)*, pages 1343–1350. International World Wide Web Conferences Steering Committee / ACM, 2013.

[29] Jérôme Kunegis and Julia Preusse. Fairness on the web: Alternatives to the power law. In *Proc. Web Science Conf.*, 2012.

[30] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, Feb 2006.

[31] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[32] Ryan N. Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 243–252, New York, NY, USA, 2010. ACM.

[33] Zhen Liu, Qian-Ming Zhang, Linyuan Lü, and Tao Zhou. Link prediction in complex networks: A local naïve bayes model. *EPL (Europhysics Letters)*, 96(4):48007, 2011.

[34] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and its Applications*, 390:1150–1170, March 2011.
[35] Linyuan Lv, Ci H. Jin, and Tao Zhou. Effective and Efficient Similarity Index for Link Prediction of Complex Networks. Technical Report arXiv:0905.3558, May 2009.
[36] Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967.
[37] Alan Mislove. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, 2009.
[38] Alan Mislove. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, Department of Computer Science, May 2009.
[39] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the Flickr social network. In *Proc. Workshop on Online Social Networks*, pages 25–30, 2008.
[40] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08)*, August 2008.
[41] Ferenc Molnar. Link Prediction Analysis in the Wikipedia Collaboration Graph, 2011.
[42] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102, Jul 2001.
[43] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.
[44] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.
[45] Andrew Chen-Brian Tran Ole J. Mengshoel, Raj Desai. Will we connect again? machine learning for link prediction in mobile social networks. 2013.
[46] Tore Opsahl and Pietro Panzarasa. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 34, 2011.
[47] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, 2001.
[48] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551, 2002.
[49] Joseph L. Rodgers and Alan W. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988.
[50] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34, 1948.
[51] Jeffrey Travers, Stanley Milgram, Jeffrey Travers, and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.
[52] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in Facebook. In *Proc. Workshop on Online Social Networks*, pages 37–42, 2009.
[53] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *In SRDS*, pages 25–34, 2003.
[54] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.
[55] Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, and Zhao Xu. Stochastic relational models for discriminative link prediction. In *Advances in Neural Information Processing Systems*, pages 333–340. MIT Press, 2007.