

## Research Article

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

**Fei Gao, Katarzyna Musial, Colin Cooper, and Sophia Tsoka**

*Department of Informatics, School of Natural and Mathematical Sciences, King's College London,  
Strand Campus, London WC2R 2LS, UK*

Correspondence should be addressed to Katarzyna Musial; katarzyna.musial@kcl.ac.uk

Received 27 February 2014; Revised 21 November 2014; Accepted 21 November 2014

Academic Editor: Jeffrey C. Carver

Copyright © 2015 Fei Gao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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, the existing methods are not comprehensively analysed. In this paper we investigate the correlation between network metrics and accuracy of different prediction methods. We selected six time-stamped real-world social networks and ten most widely used link prediction methods. The results of the experiments show that the performance of some methods has 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. Correlation analysis between network metrics and prediction accuracy of prediction methods may form the basis of a metalearning system where based on network characteristics it will be able to recommend the right prediction method for a given network.

## 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 [1]. 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 [2, 3]. 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 heterogeneous network data. During the time when ER-model was introduced, progress has also been made by sociologists in researching real-world human relationships [4, 5]. A new wave of research was set off by Watts and Strogatz who published a paper about the small-world effect in 1998 [6] and

introduction of the scale-free network model by Barabási and Albert one year later [7].

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, for example, biological networks such as protein-protein interaction networks, metabolic networks [8, 9], human networks and disease spread [10–13], scientific collaboration networks [14, 15], and online social networks [16–19].

Link prediction in complex network is one of the popular research topics. Most of the researchers focus on the link prediction problem [20] 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, and so forth (for a review of the methods please see [21]).

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; for example, 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; that is, they use classification-based methods to make predictions. Widely used methods include Decision Tree, Support Vector Machine (SVM), and Naïve Bayes [22, 23]. In [24–26], 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 systematic 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 varies 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 [27, 28] 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 predicts 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 dynamics of complex network

[29–31]. 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 in our experiments. Section 3 presents how the dataset were selected and processed. In Sections 4 and 5 we introduce the experimental design and present obtained results. We conclude the paper in Section 6.

## 2. Link Prediction Problem

Link prediction problem has been extensively studied by members of the complex network community. Liben-Nowell and Kleinberg have formalised the link prediction problem in [20] 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, that is, 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 Figure 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 (Figure 1(a)) 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 (Figure 1(b)) means that the link will disappear in the next time window. Similar to the situation when new links are added, one or more links 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 (Figure 1(c)).

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.

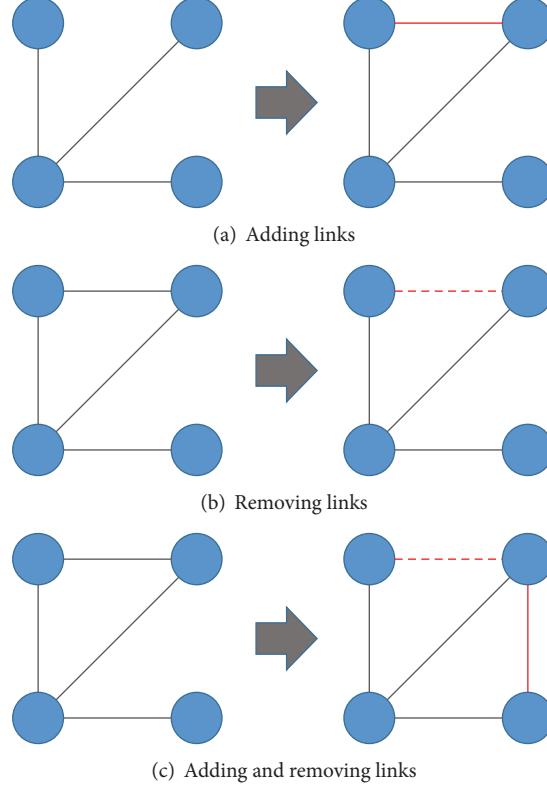


FIGURE 1: Link prediction problems.

**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 and  $k_x$  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 the 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 [17, 20, 21, 40]. The complexity of this method, as introduced in [41], is  $O(Nk^2)$ .

$$|\Gamma(x) \cap \Gamma(y)|. \quad (1)$$

**Jaccard's Coefficient.** The Jaccard 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 [20]:

$$\left| \frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)} \right|. \quad (2)$$

**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 attachment is  $O(N^2k^2)$ . This method can be expressed as

$$|\Gamma(x)| * |\Gamma(y)|. \quad (3)$$

**Adamic/Adar Index.** It was initially designed to measure the relation between personal home pages. As shown in (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 follows: 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 [16]. It is another method that is based on common neighbour; the complexity is also the  $O(Nk^2)$ . It is calculated as

$$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}, \quad (4)$$

where  $z$  is a common neighbour of node  $x$  and node  $y$ .

*Katz* <sub>$\beta$</sub> . This method takes lengths of all paths between each pair of nodes into consideration [43]. According to (5), the number of paths between node  $x$  and node  $y$  with length  $l$  (written as  $|\text{paths}_{xy}^{(l)}|$ ) is calculated and then multiplied by a factor  $\beta^l$ . By summing up all the results for the given two nodes with path length from 1 to  $\infty$ , 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  $O(N^3)$  [41, 44]:

$$\sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}_{xy}^{(l)}|. \quad (5)$$

The parameter  $\beta$ , as shown in (5), is used to adjust the weight of path with different length. When an extremely small  $\beta$  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 [21]. In network prediction problem, this method is expressed as

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

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

*Sørensen Index*. This index [45] is designed for comparing the similarity of two samples and originally used in analysis plant sociology. The complexity of this method is  $O(Nk^2)$ . It is defined as

$$\frac{2 |\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}. \quad (7)$$

*Hub Promoted Index*. HPI is proposed for analysing metabolic networks as shown in [46]. 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

$$\frac{|\Gamma(x) \cap \Gamma(y)|}{\min \{k_x, k_y\}}. \quad (8)$$

*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 the same as Hub Promoted Index,  $O(Nk^2)$ . It is defined as

$$\frac{|\Gamma(x) \cap \Gamma(y)|}{\max \{k_x, k_y\}}. \quad (9)$$

*Leicht-Holme-Newman Index*. LHNI [47] 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

$$\frac{|\Gamma(x) \cap \Gamma(y)|}{k_x * k_y}. \quad (10)$$

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 and 1 when the links between two given nodes exist). 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,  $G_{t_1}(V, L_1)$  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 nonexisting link ( $U - L_1$ ) 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 [48]. In the context of network link prediction, AUC can be interpreted as the probability that a

TABLE 1: Original dataset information.

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

This table shows the original information about the datasets used in the experiments.

<sup>a</sup>The Email network among employees of Enron. Nodes in the network are individual employees and edges are individual emails [32].

<sup>b</sup>The wall posts from the Facebook New Orleans networks [33].

<sup>c</sup>The social network of Flickr users and their friendship connections. It is collected by taking a snapshot of the network on November 2, 2006, and recording it daily until December 3, 2006, and then again daily between February 3, 2007, and May 18, 2007 [34, 35].

<sup>d</sup>The Email Communication of Wrocław University of Technology [36].

<sup>e</sup>The network contains messages sent 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 [37].

<sup>f</sup>The social network of YouTube users and their friendship connections between December 10, 2006, and January 15, 2007, and again daily between February 8, 2007, and July 23, 2007 [38, 39].

randomly chosen missing link ( $L_1 \cup L_2 - L_1$ ) is given a higher similarity score than a randomly chosen pair of unconnected links ( $U - (L_1 \cup L_2)$ ) [49]. The algorithmic implementation of AUC follows the approach in [21]. It is calculated as

$$\frac{n' + 0.5n''}{n}, \quad (11)$$

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 is generated from an independent and identical distribution. Thus, the degree to which the AUC exceeds 0.5 indicates how much better the predictions are when compared to prediction 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 [50]) and another one 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 was 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 are selected due to the calculation capacity. As for some dense networks, 8000 nodes are also too big, so we choose 6000) from the original dataset as the sample user data. As UC Irvine Messages only contain 1899 users, so we leave them as they are. 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 of 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 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 both sets have the same nodes  $V$ .

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

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 the table.

## 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 unweighted network. This enables us to reach a 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 is presented in [21] 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 [50] and MIT Strategic Engineering research group. The metrics we calculate include the following.

*Global Clustering Coefficient.* It is defined in [51] as

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

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

*Average Clustering Coefficient* [6]. It is based on local clustering  $C_l$ . For each vertex  $l$ , its local clustering coefficient can be calculated by

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

and then the ACC can be calculated as

$$\text{ACC} = \frac{1}{v} \sum_l C_l, \quad (14)$$

where  $v$  is the number of nodes in a network.

*Network Density.* The ratio between number of existing links and number all possible links within a given network.

$$\text{Network Density} = \frac{\text{Number of Existing Links}}{\text{Number of all possible links}}, \quad (15)$$

where

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

where  $v$  is the number of nodes in the network.

*Gini Coefficient* [52]. In the network theory Gini coefficient is defined as

$$G = \frac{2 \sum_{i=1}^n i d_i}{n \sum_{i=1}^n d_i} - \frac{n+1}{n}, \quad (17)$$

where  $d_1 \leq d_2 \leq d_3 \leq \dots \leq d_n$  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.* It is the longest path out of the set of all shortest paths in the network.

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

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 is

$$\text{ASP} = \frac{1}{v \cdot (v - 1)} \cdot \sum_{i \neq j} d(i, j). \quad (19)$$

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 [53] 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

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

TABLE 3: Theoretical GCC and 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

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.

## 5. Experiment Result

**5.1. Network Profiles.** The values of network metrics for each of the extracted social networks 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 paths in our experiments are all smaller than six, the number suggested by the six degrees of separation theory [54]. The average shortest path of the six 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 [55] where authors established that the average shortest path of Twitter is 3.43.

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

TABLE 4: Analytical formulas for GCC and ASP in random and regular networks.

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

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

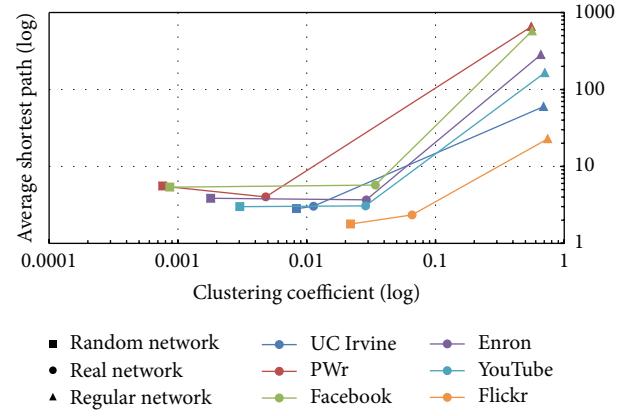


FIGURE 2: GCC and ASP calculated for six analysed networks (circle) and corresponding random (square) and regular (triangle) networks generated using the same number of nodes and connections as in real-world networks. Different colours depict different networks and corresponding random and regular networks—see legend. For example green circle is a datapoint for Facebook network, green square is a datapoint for random network corresponding to Facebook network and green triangle is a datapoint for regular network corresponding to Facebook network.

We also compared the GCC and ASP metrics of the real network with the theoretical metrics of random network and regular network that have the 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.

Figure 2 plots the metrics of six analysed networks and related theoretical networks, respectively. It shows that the clustering coefficients of the analysed networks are all between random and regular networks. Meanwhile, the average shortest paths 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

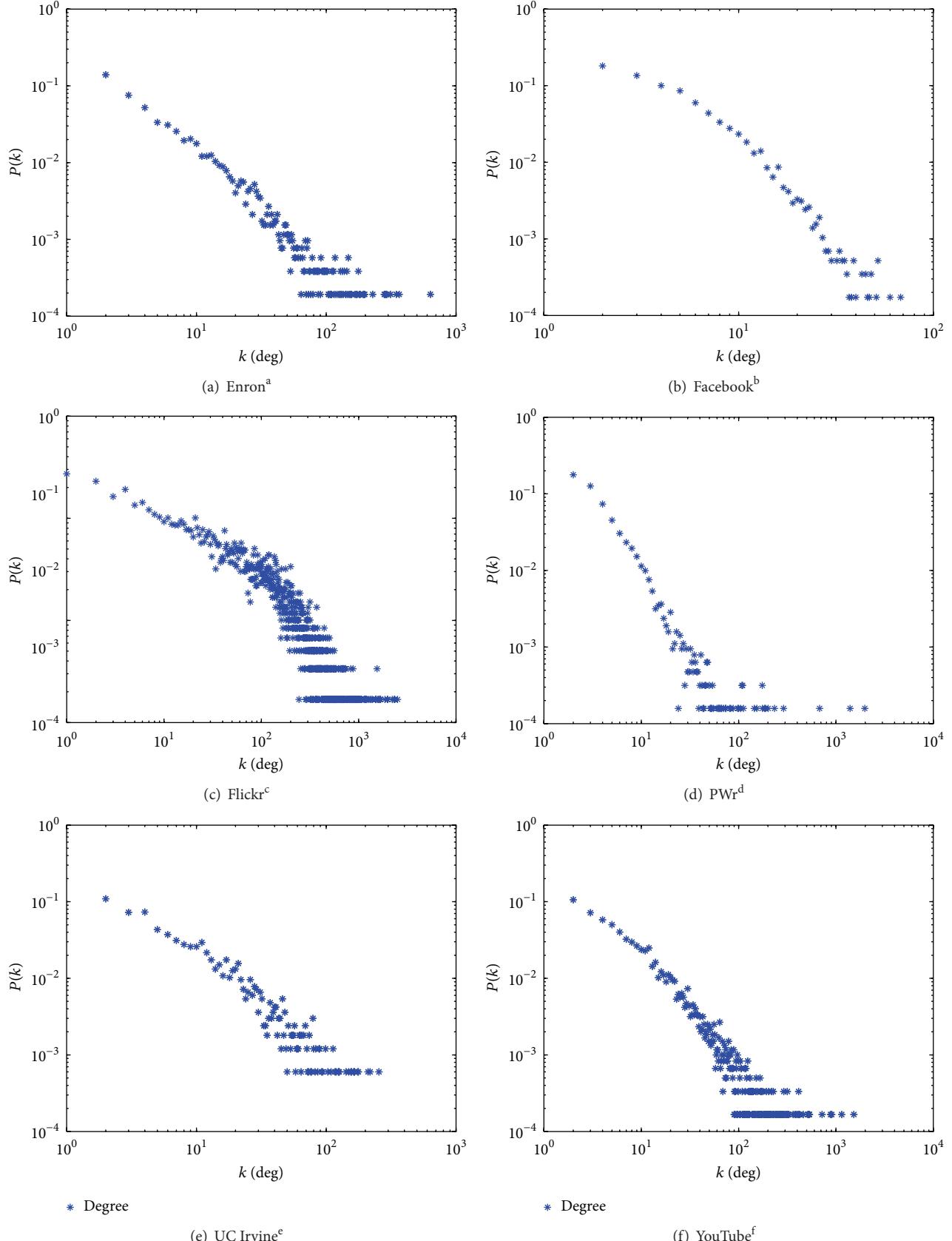


FIGURE 3: The Degree distributions. The degree distributions are all following the power law with exponent <sup>a</sup>Enron,  $r = 1.85$ ; <sup>b</sup>Facebook,  $r = 1.82$ ; <sup>c</sup>Flickr,  $r = 1.25$ ; <sup>d</sup>PWr,  $r = 2.19$ ; <sup>e</sup>UC Irvine,  $r = 1.56$ ; <sup>f</sup>YouTube,  $r = 1.56$ .

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

TABLE 6: Prediction methods accuracy result (AUC).

Datasets	AUC									
	CN	JI	PA	AA	Katz <sub><math>\beta</math></sub> <sup>a</sup>	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

The accuracy of selected prediction methods measured by AUC. The average performance and the variance for each method are also listed.

<sup>a</sup>In our experiment, we choose  $\beta = 0.0005$ .

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 relative 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 networks “prediction unfriendly” networks. Please note that, in the experiments, for both prediction friendly and unfriendly networks, Katz <sub>$\beta$</sub>  always provides a good performance level.

**5.3. Correlation between Prediction Accuracy and Network Metrics.** Table 7 shows 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. 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.

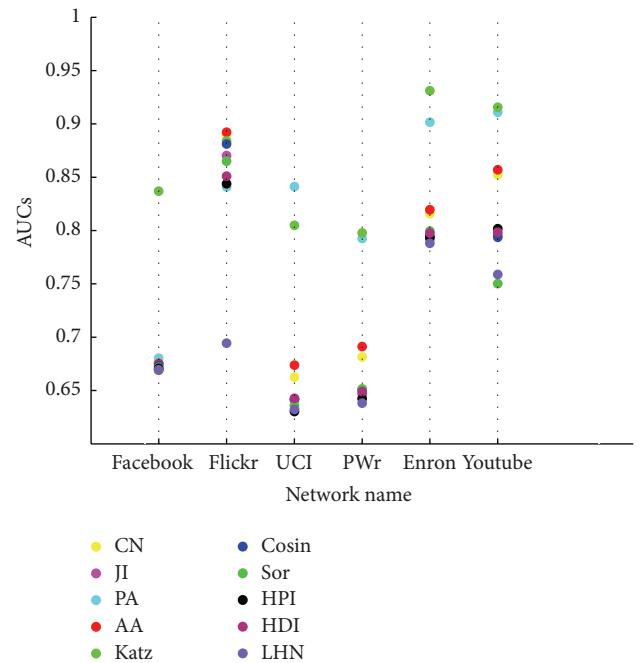


FIGURE 4: The AUC prediction results for each network.

In Figure 5, we can see that the preferential attachment and Gini Coefficient provide 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

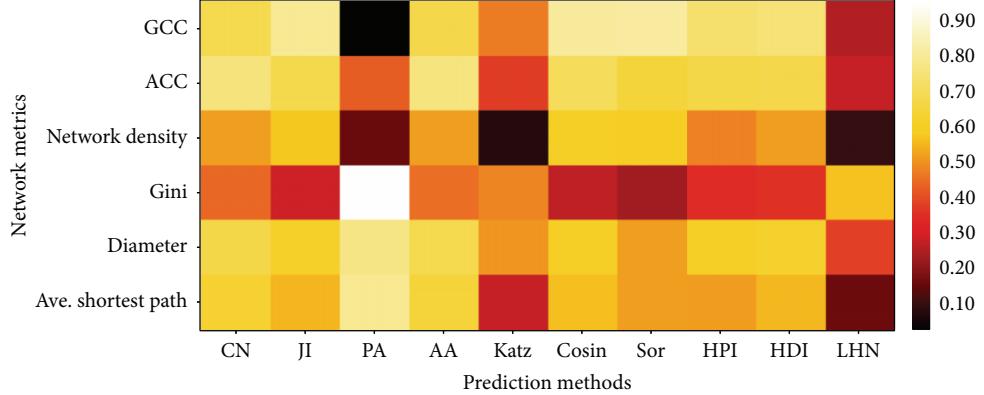


FIGURE 5: Heat-map of network metrics and prediction methods correlation. As for the Pearson Coefficient, both 1 and -1 stand for linear relationship (positive and negative); we use the absolute value of correlation coefficient in this figure to indicate whether the two factors are linearly correlated.

TABLE 7: Pearson Correlation of prediction methods accuracy and network metrics.

	CN	JI	PA	AA	Katz <sub><math>\beta</math></sub>	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

This table shows the correlation between prediction methods accuracy and network metrics calculated with Pearson's linear correlation coefficient. The number is 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

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 they do in networks with 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 coefficient. Based on the above we can say that prediction methods positively correlated with GCC and negatively with ASP and diameter 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 the 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 of 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, and 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.

## 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 <sub>$\beta$</sub> . 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.

The Pearson correlation coefficient enabled us 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, S-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 combines 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.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

- [1] N. L. Biggs, E. K. Lloyd, and R. J. Wilson, *Graph Theory*, New York, NY, USA, The Clarendon Press, 2nd edition, 1986.
- [2] P. Erdős and A. Rényi, “On random graphs. I,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.
- [3] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.
- [4] J. Travers and S. Milgram, “An experimental study of the small world problem,” *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.
- [5] S. Milgram, “The small world problem,” *Psychology Today*, vol. 2, pp. 60–67, 1967.
- [6] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [7] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [8] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
- [9] A. D. King, N. Pržulj, and I. Jurisica, “Protein complex prediction via cost-based clustering,” *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.
- [10] R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, 2001.
- [11] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, “Epidemic spreading in real networks: an eigenvalue viewpoint,” in *Proceedings of the 22nd International Symposium on Reliable Distributed Systems (SRDS ’03)*, pp. 25–34, October 2003.
- [12] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, “Epidemic thresholds in real networks,” *ACM Transactions on Information and System Security*, vol. 10, no. 4, article 1, 26 pages, 2008.
- [13] N. A. Christakis and J. H. Fowler, “The spread of obesity in a large social network over 32 years,” *The New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.
- [14] Z. Huang, X. Li, and H. Chen, “Link prediction approach to collaborative filtering,” in *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL ’05)*, pp. 141–142, ACM, New York, NY, USA, June 2005.
- [15] F. Molnar, “Link prediction analysis in the Wikipedia Collaboration graph,” 2011, <http://www.cs.rpi.edu/~magdon/courses/casp/projects/Molnar.pdf>.
- [16] L. A. Adamic and E. Adar, “Friends and neighbors on the Web,” *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

- [17] W. Cukierski, B. Hamner, and B. Yang, "Graph-based features for supervised link prediction," in *Proceedings of the International Joint Conference on Neural Network (IJCNN '11)*, pp. 1237–1244, August 2011.
- [18] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT '11) and IEEE International Conference on Social Computing (SocialCom '11)*, pp. 73–80, Boston, Mass, USA, October 2011.
- [19] K. Juszczyszyn, K. Musiał, and M. Budka, "Link prediction based on Subgraph evolution in dynamic social networks," in *Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT '11) and IEEE International Conference on Social Computing (SocialCom '11)*, pp. 27–34, October 2011.
- [20] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM '03)*, pp. 556–559, ACM, New York, NY, USA, November 2003.
- [21] L. Lü and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and Its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
- [22] O. J. Mengshoel, R. Desai, A. Chen, and B. Tran, "Will we connect again? Machine learning for link prediction in mobile social networks," 2013.
- [23] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: a local naïve Bayes model," *Europhysics Letters*, vol. 96, no. 4, Article ID 48007, 2011.
- [24] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 243–252, New York, NY, USA, July 2010.
- [25] K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu, "Stochastic relational models for discriminative link prediction," in *Advances in Neural Information Processing Systems*, pp. 333–340, MIT Press, Boston, Mass, USA, 2007.
- [26] M. Al Hasan, V. Chaoji, S. Salem, and Z. Mohammed, "Link prediction using supervised learning," in *Proceedings of the SDM 6th workshop on Link Analysis, Counterterrorism and Security*, 2006.
- [27] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.
- [28] P. Aloy and R. B. Russell, "InterPreTS: protein interaction prediction through tertiary structure," *Bioinformatics*, vol. 19, no. 1, pp. 161–162, 2003.
- [29] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: structure and dynamics," *Physics Reports: A Review Section of Physics Letters*, vol. 424, no. 4–5, pp. 175–308, 2006.
- [30] M. Budka, K. Juszczyszyn, K. Musiał, and A. Musiał, "Molecular model of dynamic social network based on e-mail communication," *Social Network Analysis and Mining*, vol. 3, no. 3, pp. 543–563, 2013.
- [31] C. Guido, A. Chessa, I. Crimaldi, and F. Pammolli, *The Evolution of Complex Networks: A New Framework*, 2012.
- [32] B. Klimt and Y. Yang, "The Enron corpus: a new dataset for email classification research," in *Proceedings of the 15th European Conference on Machine Learning (ECML '04)*, pp. 217–226, September 2004.
- [33] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Online Social Networks*, pp. 37–42, Barcelona, Spain, August 2009.
- [34] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the Flickr social network," in *Proceedings of the Workshop on Online Social Networks*, pp. 25–30, 2008.
- [35] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN '08)*, pp. 25–30, ACM, August 2008.
- [36] P. Kazienko, K. Musiał, and A. Zgrzywa, "Evaluation of node position based on email communication," *Control and Cybernetics*, vol. 38, no. 1, pp. 67–86, 2009.
- [37] T. Opsahl, "Triadic closure in two-mode networks: redefining the global and local clustering coefficients," *Social Networks*, vol. 35, no. 2, pp. 159–167, 2013.
- [38] A. Mislove, *Online social networks: measurement, analysis, and applications to distributed information systems [Ph.D. thesis]*, Rice University, 2009.
- [39] A. Mislove, *Online social networks: measurement, analysis, and applications to distributed information systems [Ph.D. thesis]*, Department of Computer Science, Rice University, 2009.
- [40] H. R. de Sa and R. B. C. Prudencio, "Supervised link prediction in weighted networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '11)*, pp. 2281–2288, August 2011.
- [41] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, Article ID 046122, 2009.
- [42] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, Article ID 025102, 2001.
- [43] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.
- [44] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.
- [45] T. Sørensen, *A Method of Establishing Groups of Equal Amplitude in Plants Ociology Based on Similarity of Species and Its Application to Analyses of The Vegetation on Danish Commons*, vol. 5 of *Biologiske Skrifter*, E. Munksgaard, 1948.
- [46] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.
- [47] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, no. 2, Article ID 026120, 2006.
- [48] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [49] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.
- [50] J. Kunegis, "KONECT—the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, pp. 1343–1350, May 2013.

- [51] M. E. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
- [52] J. Kunegis and J. Preusse, "Fairness on the Web: Alternatives to the power law," in *Proceedings of the 3rd Annual ACM Web Science Conference (WebSci '12)*, pp. 175–184, June 2012.
- [53] J. L. Rodgers and A. W. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.
- [54] M. E. J. Newman, A. L. Barabási, and D. J. Watts, Eds., *The Structure and Dynamics of Networks*, Princeton Studies in Complexity, Princeton University Press, Princeton, NJ, USA, 2006.
- [55] R. Bakhshandeh, M. Samadi, Z. Azimifar, and J. Schaeffer, "Degrees of separation in social networks," in *Proceedings of the 4th International Symposium on Combinatorial Search (SoCS '11)*, pp. 18–23, July 2011.

