Scientometrics (2018) 117:381–390
https://doi.org/10.1007/s11192-018-2867-7

h-Index-based link prediction methods in citation network

Wen Zhou[1,2] · Jiayi Gu[1] · Yifan Jia[1]

Received: 23 February 2018 / Published online: 3 August 2018

Abstract
Link prediction implies the mining of the missing links in networks or prediction of the next node pair to be connected by a link. Link prediction is useful for mining information in citation networks, and most of the existing related studies commonly use degree rather than more advanced methods to measure the importance of nodes. However, such a method cannot easily measure the importance of a paper in reality; some papers have high degree in citation networks but are not very influential. This issue restricts the performance of the link prediction methods applied to citation networks. The current study analyzed *h*-type indices, which are more suitable than degree for measuring the importance of citation network nodes. We propose two *h*-index-based link prediction methods. Experiments conducted on real citation networks demonstrate that the use of *h*-type index to measure the importance of nodes in citation networks can significantly improve the prediction accuracy of link prediction methods.

Keywords Complex network · Link prediction · *h*-Index · Citation network · Graph mining

Introduction

Research requires a search for a large number of academic papers. Citation networks, which are networks of citations among academic papers, are widely used in many applications, such as citation recommendation, discovery of research hotspot, and finding experts. In some of these applications researchers always want to know which suitable paper to cite in the future, that is, the next status of a node or the next node pair to be connected by a link. Link prediction in complex networks is committed to estimating the likelihood of the existence of a link between two nodes based on the topological structure of a network and the nodal attributes (Getoor and Diehl 2005).

Similarity-based algorithms are the commonly used link prediction methods, in which each pair of nodes possesses a so-called proximity score. Proximity score can be defined

✉ Wen Zhou
zhouwen@shu.edu.cn; wen.zhou@ri.se

✉ Yifan Jia
asdfzjyf@163.com

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

[2] RISE SICS, 16440 Stockholm, Sweden

based solely on the network structure, implying that two nodes are considered to be similar if they have a high structural similarity.

The existing set of similarity-based algorithms may be organized into several different categories. Local structural similarity methods only investigate the local topological information in a network. The simplest local structural similarity method is the Common-Neighbors (CN) index. In general, the more common neighbors the two nodes have, the more likely they are to have a link. Based on the CN index, degree is considered to improve methods, thus leading to the generation of other local structural similarity methods, including Salton, Sorenson, Jaccard coefficient, Adamic–Adar (AA), and Leicht–Holme–Newman (LHN1) indices. Global structural similarity methods consider the structure of the entire network, such as Katz, LHN2, Average Commute Time, Random Walk with Restart, SimRank and Matrix Forest Index. The introduction and comparison between these similarity indices have been presented by Lü and Zhou (2011).

Moreover, some methods based on maximum likelihood estimation were proposed (Clauset et al. 2008; Guimerà and Sales-Pardo 2009). In all the above-mentioned methods, local structural similarity methods are more commonly used because of their lower computing complexity. Hence, the current paper focuses on the methods of local similarity indices.

In recent years, some great accomplishments have been made in the field of complex network research. The research on measuring the importance of a node has attracted considerable attention (Gleich 2015; Ermann et al. 2015) because of its great theoretical significance as well as a wide range of applications. There are several methods to identify an important node, such as k-core (Yang et al. 2017). The evaluation of the importance of a node greatly affects the accuracy of link prediction methods. However, the existing link prediction methods commonly use degree to measure the importance of a node. In different real world networks, a node represents several different entities and its importance cannot be measured through one standard. In citation networks, some papers with high degree may not be highly influential. There are a number of factors to determine whether a paper is important. Schubert suggested an extension of the *h*-index concept to assess the influence of a paper (Schubert 2009). In this paper, we propose two new *h*-index-based link prediction methods: *h*-Salton and *h*-AA. For these methods, *h*-index is chosen as an advanced measure for determining the importance of a node in a citation network. The *h*-index we utilized in this study is called lobby index, which is a version of Schubert's *h*-index (Korn et al. 2009; Schubert 2009). We tested these two methods on real citation networks, and the results demonstrated that the use of *h*-index to measure the importance of nodes can achieve better performance at a higher link-prediction accuracy.

The remainder of this paper is structured as follows. "Materials and methods" section introduces the data and new methods. "Experiments and empirical analysis" section shows the experimental results, and "Discussion and conclusion" section draws the study conclusion.

Materials and methods

First, this section first introduces the networks used for testing our methods. Second, a brief introduction of citation networks and *h*-index is presented. Then, the *h*-index-based methods proposed to improve the existing methods are described. Finally, we state an evaluation metric to evaluate the performance of the new methods.

Data description

The following three citation networks were used in this study.

- Small & Griffith and Descendants (SmaGri) (Batagelj and 2006). This is the network of citations to Small & Griffith and Descendants. (See Pajek Datasets: http://vlado.fmf.uni-lj.si/pub/networ ks/data/cite/SmaGri.zip.)
- Kohonen (Batagelj and Mrvar 2006). This is the network of articles with topic ''self-organizing maps'' or references to ''Kohonen T''. (See Pajek Datasets: http://vlado.fmf.uni-lj.si/pub/networks/data /cite/Kohonen.zip.)
- Scientometrics (SciMet) (Batagelj and Mrvar 2006). This is the network of articles from or citing Scientometrics. (See Pajek Datasets: http://vlado.fmf.uni-lj.si/pub/networks/data/cite/Sci Met.zip.)

Table 1 summarizes the basic topological measures of these networks. If a network is unconnected, we only consider its largest weakly connected component.

Citation networks and *h*-index

The network of citations between academic papers is a kind of classic information network. In citation networks, nodes represent academic papers, and a directed link exists from paper A to B if A cited B in its references. Citation networks are similar to the World Wide Web; however, at least one important difference exists between the two: a citation network is acyclic, while the Web is not. An acyclic network is one in which there are no closed loops of directed edges (Newman 2010). Citation networks are acyclic because a paper can only cite an already written paper. Moreover, citation networks have some statistics, that is, approximately 47% of all papers in the Science Citation Index have never been cited at all; of the remainder, 9% have one citation, 6% have two, after which these statistics decrease rapidly. Only 21% of all papers have 10 or more citations, and only 1% have 100 or more. These figures show the characteristics of the power-law distribution in citation networks (Newman 2010; Price 1965).

In addition, a common method to measure the impact of a paper in citation networks is to calculate its degree, that is, the number of citations a paper receives (Uzzi and Jones 2013). The existing link prediction methods, such as Salton index, Sorenson index and AA index, usually use degree to measure the importance of a node. Kitsak et al. (2010) argued that the degree cannot describe the importance of a node in a complex network completely. According to the statistics of citation networks mentioned earlier, a large number of papers are not cited or are rarely cited by other papers; thus, these papers may not be of much

Table 1 Basic topological measures of three real networks

Nets	$\|V\|$	$\|E\|$	$\langle k \rangle$	$\langle L \rangle$	C
SmaGri	1024	4919	4.8	3.242	0.302
Kohonen	3704	12,683	3.4	3.272	0.252
SciMet	2678	10,381	3.9	4.229	0.174

$|V|$ and $|E|$ are the number of nodes and links respectively, $\langle k \rangle$ is the average degree of all nodes, $\langle L \rangle$ is the average path length, C represents the clustering coefficient, which is defined as the number of closed triplets divided by the number of connected triplets of nodes

value, and papers cited by these papers may also not be valuable. That is, a node pointed by many valueless nodes may lead to the incorrect conclusion that the node is a vital node.

h-Index also called the Hirsch index is greatly used to evaluate the academic achievement of a scholar (Ayaz et al. 2017). It is defined as the highest number of papers of a scholar that have been cited no less than *h* (Hirsch 2005). Lobby index as an *h*-type index is an extension of *h*-index and refers to individual nodes as the node centrality measure in networks (Korn et al. 2009; Campiteli et al. 2013; Lü et al. 2016). In 2009, a degree *h*-index was proposed as an indicator for complete networks (Schubert et al. 2009). Later, Glänzel mentioned that core documents can be defined on the basis of a network's degree *h*-index (Glänzel 2012). In our method, we utilize an *h*-index extension called lobby index in order to measure the importance of papers in a citation network. Here the *h*-index of a node in a network can be defined as follows: the in-degree of an arbitrary node i is denoted by k_i. A_{ij} is the adjacency matrix, $A_{ij} \in [0, 1]$. $A_{ij}\ k_j$ is nonzero only when paper j cites paper i, being equal to k_j. Then, we construct an operator H, which acts on a finite number of reals $(x_1, x_2, \ldots, x_n)$ and returns integer $y = H(x_1, x_2, \ldots, x_n) > 0$, where y is the maximum integer such that there exist at least y elements in $(x_1, x_2, \ldots, x_n)$, each of which is no less than y. Then, the *h*-index of node i is given as function (1).

$$h_i = H\left(A_{i1}k_1, \ldots, A_{ij}k_j, \ldots, A_{in}k_n\right), \tag{1}$$

where $j = 1, \ldots, n$ is the index of the paper and n is the total number of nodes in the network. $A_{ij}\ k_j$ is the degree of node i's neighbors.

Compared with degree, *h*-index was determined to strengthen the importance of nodes with several high in-degree neighbors and decrease the importance otherwise. As shown in Fig. 1, node i_1 has a degree of 5, which is higher than the degree of 3 of node i_2. However, according to its definition, *h*-index of i_1 is 0, which is lower than 3 of i_2. This is much closer to our understanding of a citation network because although the in-degree of i_2 is not higher than that of i_1, its neighbors pointing to it have a higher degree than that of i_1. Thus, the importance of i_2 in the network should be greater. Moreover, the neighbors of i_1 have no in-degree. This kind of node is either inactive or unimportant, and is thus dispensable in the network.

To describe the difference between degree and *h*-index more intuitively, we consider the previously mentioned three real-world citation networks as examples. The scatter diagrams of the node degrees of each network and *h*-index are shown in Figs. 2, 3 and 4.

The scatter diagrams show that the distribution of a node's degree values is broader than the distribution of its *h*-index values. Some nodes with a higher degree are on the same *h*-

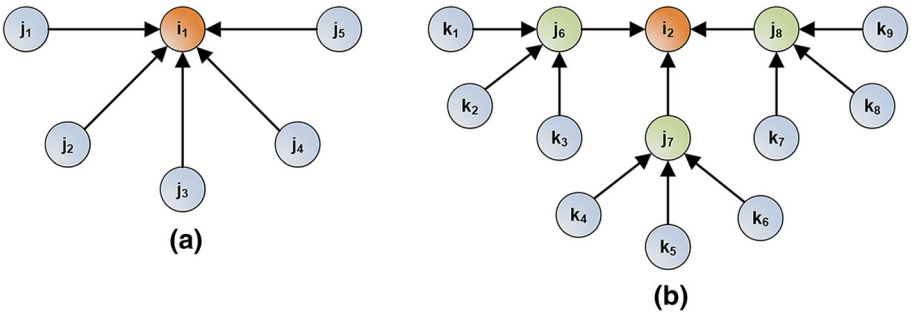

Fig. 1 Illustration of the difference between degree and *h*-index

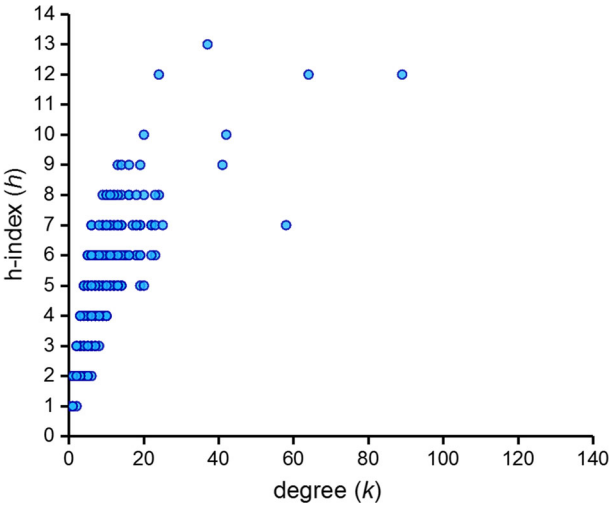

Fig. 2 Scatter diagram of degree of nodes and *h*-index of SmaGri network

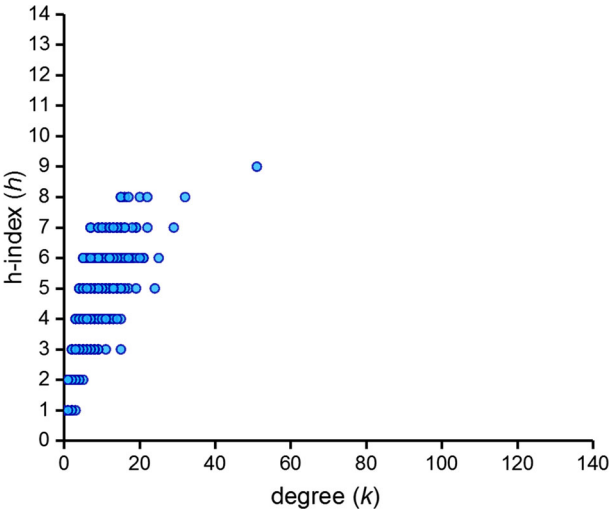

Fig. 3 Scatter diagram of degree of nodes and *h*-index of Kohonen network

index level as the nodes with a lower degree. Therefore, a considerable difference exists between the use of degree and *h*-index to measure the importance of nodes. However, existing link prediction methods commonly use degree to evaluate the importance of nodes. *h*-index is also a good node evaluation method, and Schubert believed this evaluation method can assess the influence of a paper (Schubert 2009).

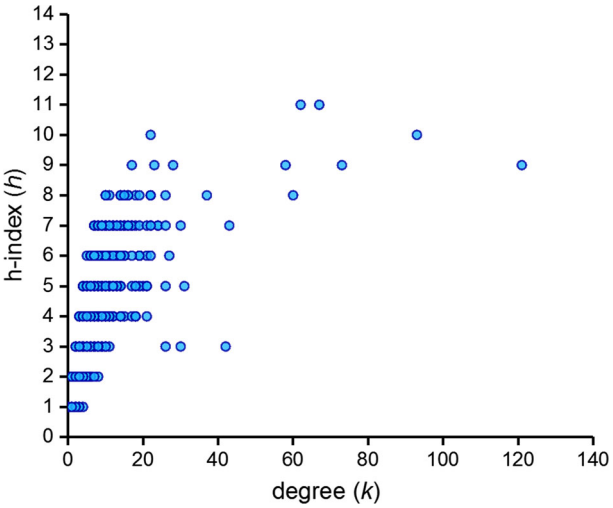

Fig. 4 Scatter diagram of degree of nodes and *h*-index of SciMet network

h-Salton and *h*-AA methods

This study focused on two link prediction methods using *h*-index, that is, *h*-Salton and *h*-AA, to measure the importance of nodes in citation networks. These two methods are variants of the Salton and AA indices, respectively, which replace degree with *h*-index.

Consider an un-weighted, directed simple network given by $G(V, E)$, where V is the set of nodes and E is the set of links. Multiple links and self-connections were not considered in this study, $\mathbf{A}$ denotes the network's adjacency matrix. For each pair of nodes $x, y \in V$, $\mathbf{A}_{xy}= 1$ if x and y are linked and $\mathbf{A}_{xy}= 0$ otherwise. We directly set the existence likelihood of link (x, y) as score $\mathbf{S}_{xy}$ to each potential link; a higher score implies a higher likelihood that a link exists. In some link prediction methods, the scores may not be directly related to a certain measurement but describe the existence likelihood of links. In such cases, all unlinked pair of nodes will be ranked in the descending order of their scores so that the links on the top of the rank can be considered to have the highest existence likelihoods. Each score $\mathbf{S}_{xy}$ can be used to constitute similarity matrix $\mathbf{S}$, which is used as the evaluation metric to measure the prediction accuracy.

The definitions of *h*-Salton and *h*-AA are as follows:

h-Salton index

The Salton index (Salton and McGill 1986) is also called the cosine similarity in literature, and is defined as function (2):

$$S_{xy}^{\text{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}}, \tag{2}$$

where k_x is the degree of node x and $\Gamma(x)$ is the set of neighbors of node x.

The *h*-Salton Index is the variant of the Salton index that replaces degree with *h*-index, and is defined as function (3):

$$S_{xy}^{h-\text{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{h_x \times h_y}}, \tag{3}$$

where h_x is the h-index of node x.

h-AA index

The AA index (Adamic and Adar 2003) is one variant of the CN index that assigns the less connected neighbors more weights. Its definition is as function (4):

$$S_{xy}^{AA} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}. \tag{4}$$

The h-AA Index is the variant of the AA index that replaces degree with h-index, and is defined as function (5):

$$S_{xy}^{h-AA} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log h_z}. \tag{5}$$

Evaluation metric

To evaluate the h-Salton and h-AA methods, this study compared their prediction accuracies with those of the Salton and AA indices, respectively. For each network $G(V, E)$, link set E is randomly divided into training set E_T and test set E_V, such that $E_T \cup E_V = E$ and $E_T \cap E_V = \varnothing$. A nonexistent link is defined as the link in $U - E$, where U is a universal set for $(|V| \times (|V| - 1)) / 2$ links and $|V|$ is the number of elements in set V. The link in the test set is called a missing link. The task of a link prediction method is to find missing links by using the training set. Through a link prediction method, each missing and nonexistent link is assigned a proximity score, which represents the likelihood of the existence of links. The area under the receiver operating characteristic (ROC) curve (AUC) was used to measure the prediction accuracy. AUC can be defined as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. After providing a proximity score to each missing and nonexistent link, we randomly selected a missing and a nonexistent link and compared their scores to compute the AUC value. If among n independent comparisons, the missing links have a higher score n' times and the same score n'' times, the AUC value is defined as function (6):

$$\text{AUC} = \frac{n' + 0.5n''}{n}. \tag{6}$$

If all missing links have the same score as the nonexistent links, the probability of the method finding all the missing links is 0.5, and $n'' = n$ and $n' = 0$. The AUC value should be approximately 0.5 if all the scores are generated from an independent and identical distribution (Lü and Zhou 2011). If all the missing links have a higher score compared to the nonexistent links, the method correctly finds all the missing links, and $n' = n$ and $n'' = 0$. In addition, the AUC value will take the highest possible value of 1. Therefore, the AUC value of a good link prediction method should be close to 1.

Experiments and empirical analysis

In this study, four indices were applied on three empirical citation networks. As the link set is randomly divided into a training set and a test set, and the selection of a missing link and a nonexistent link for comparison is random, we experimented multiple times for each AUC value. For each network, the link set was divided five times, and for each division, the AUC value was computed four times. Therefore, each AUC value was obtained by averaging 20 independent experiments. To observe the effect of link set E being split by different proportions for the AUC value, three sizes of training sets were used. The experimental results are shown in Figs. 5, 6 and 7. As observed the AUC values of the h-Salton and h-AA methods are higher than the original values, and the effect of replacing degree with h-index on the Salton index is better than that on the AA index.

The Salton and AA indices as local similarity indices are different. They utilize different structural features to pursue a high prediction accuracy, and show different performance on the same network. To estimate the likelihood of the existence of a directed link between two papers in a citation network, the Salton index considers the influence of the number of common neighbors between the two papers and the degree of each paper (the h-index instead of the degree of each paper is used in the h-Salton index); the AA index considers the degree of common neighbors (the h-index of common neighbors between the two papers is used in the h-AA index). Thus, the improvement of replacing degree with h-index on the Salton and AA indices differs.

Discussion and conclusion

In this paper, we proposed h-index is more suitable for measuring the importance of citation network nodes through the empirical analysis of citation network characteristics. Second to prove that the methods of measuring important nodes more suitable for network characteristics have a positive effect on the application of link prediction methods to real networks, we replaced degree with h-index in the Salton and AA indices, and proposed two link prediction methods h-Salton and h-AA, respectively. Finally, we used AUC to measure the prediction accuracy and conduct experiments on real citation networks, and demonstrated that the use of h-index instead of degree can cause the link prediction

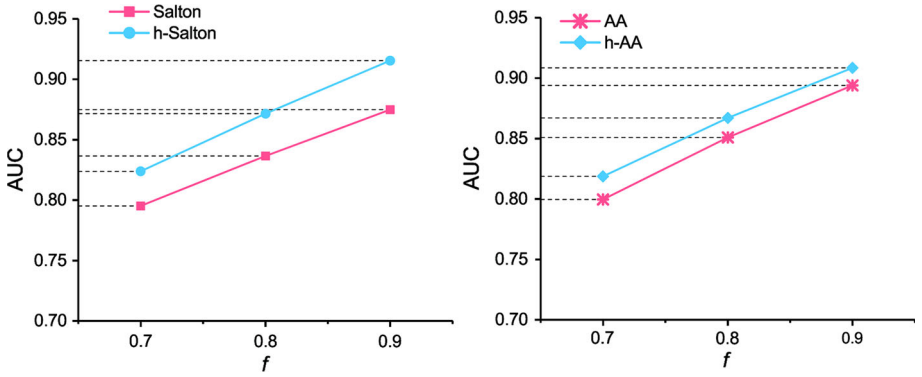

Fig. 5 AUC values of different indices are applied on SmaGri network with different sizes of training sets (f symbolizes the fraction of links in the training set)

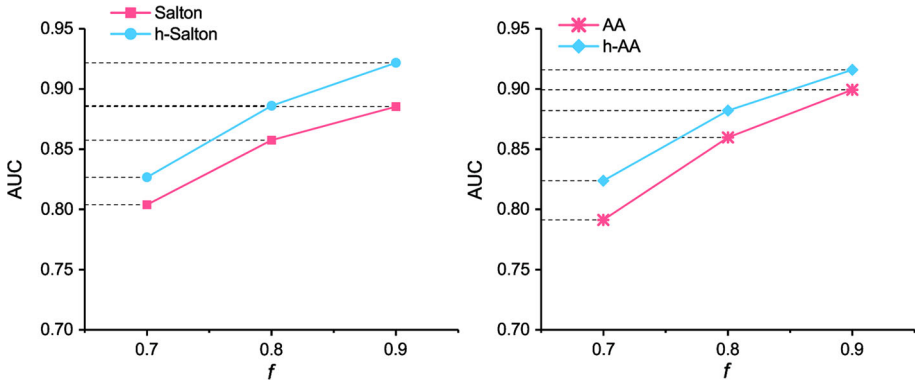

Fig. 6 AUC values of different indices are applied on Kohonen network with different sizes of training sets (*f* symbolizes the fraction of links in the training set)

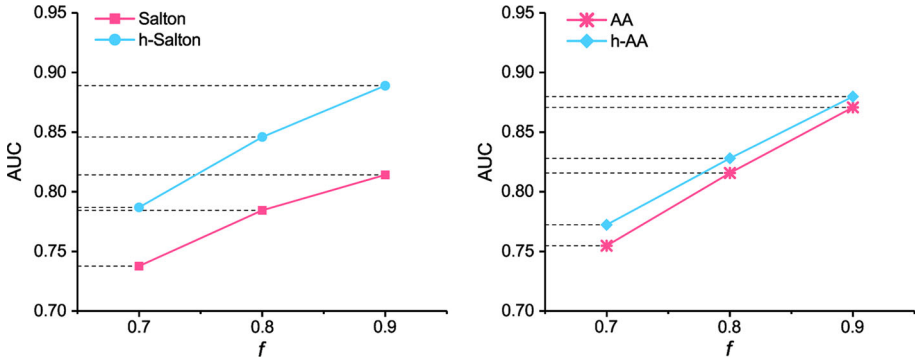

Fig. 7 AUC values of different indices are applied on SciMet network with different sizes of training sets (*f* symbolizes the fraction of links in the training set)

methods to show a higher prediction accuracy in citation networks. In addition, we discussed why the effect on the *h*-Salton index versus the Salton index is better than that on the *h*-AA index versus the AA index.

This study was aimed at using a new method to measure the importance of a node in link prediction methods. A good result was obtained, thereby supporting our idea. The preliminary results show the improvement of link prediction methods and their application in citation networks. However, more detailed research is needed in the future to determine whether *h*-index is also suitable for other kinds of networks, and not just citation networks, and whether other existing node evaluation methods could improve the performance of the Salton and AA indices.

Acknowledgements This work is supported by National Natural Science Foundation of China (NSFC No. 71203135).

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 25*(3), 211–230.

Ayaz, S., Masood, N., & Islam, M. A. (2017). Predicting scientific impact based on h-index. *Scientometrics, 114*(3), 993–1010.

Batagelj, V., & Mrvar, A. (2006). Pajek datasets website. http://vlado.fmf.uni-lj.si/pub/networks/data/. Accessed 14 Jan 2013.

Campiteli, M. G., Holanda, A. J., Soares, L. D. H., Soles, P. R. C., & Kinouchi, O. (2013). Lobby index as a network centrality measure. *Physica A: Statistical Mechanics and Its Applications, 392*(21), 5511–5515.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature, 453*(7191), 98–101.

Ermann, L., Frahm, K. M., & Shepelyansky, D. L. (2015). Google matrix analysis of directed networks. *Reviews of Modern Physics, 87*(4), 1261.

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter, 7*(2), 3–12.

Glänzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics, 93*(1), 113–123.

Gleich, D. F. (2015). PageRank beyond the Web. *SIAM Review, 57*(3), 321–363.

Guimerà, R., & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences, 106*(52), 22073–22078.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., et al. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893.

Korn, A., Schubert, A., & Telcs, A. (2009). Lobby index in networks. *Physica A: Statistical Mechanics and Its Applications, 388*(11), 2221–2226.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications, 390*(6), 1150–1170.

Lü, L., Zhou, T., Zhang, Q.-M., & Stanley, H. E. (2016). The H-index of a network node and its relation to degree and coreness. *Nature Communications, 7,* 10168.

Price, D. J. (1965). Network of scientific papers. *Science, 149*(3683), 510–515.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York City: McGraw-Hill Inc.

Schubert, A. (2009). Using the h-index for assessing single publications. *Scientometrics, 78*(3), 559–565.

Schubert, A., Korn, A., & Telcs, A. (2009). Hirsch-type indices for characterizing networks. *Scientometrics, 78*(2), 375–382.

Uzzi, B., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.

Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Yang, Y., Nishikawa, T., & Motter, A. E. (2017). Small vulnerable sets determine large network cascades in power grids. *Science, 358*(6365), eaan3184.