An improved degree measurement index based on the social network

Lu Dai*, Guangming Li

Colledge of computer, Dongguan University of Technology, Dongguan 523808, Dongguan

**Corresponding author's e-mail: dailudgut@163.com*

Received 01 December 2013, www.cmnt.lv

Abstract

This paper stems from discussion of the structure of the scientific collaboration network, utilizes t methods of complex network analysis and social network analysis, give out an analysis on the scientific collaboration network from both macro and micro perspectives, and in both dynamic and static way. This paper extracts 16 years of conference proceedings articles from 1995 to 2014 as an experimental data set, with the corresponding network known as collaboration network of data mining. Based on analysis of the classic centrality metric, an improved node centrality metrics (c-index) is proposed, to measure collaboration strength of nodes in a weighted collaboration network.

Keywords: social network, collaboration network, centrality measurement

1 Introduction

In essence, the cooperation strength and the degree of a node (the number of cooperators) and the edge's strength (the cooperation frequency) are related with the importance of the neighbor nodes (cooperators). So how should we take these indexes into comprehensive consideration? We shall improve the index by combining H-index with social network.

Firstly, we shall introduce the H-index. It is a method for measuring academic achievements or the scholars' influence by combining the number of being cited with the number of articles, which was proposed by Hirsh [1] in 2005. For instance, if one scientist owns NP (the number of articles) articles, and each of them have been cited for h times at least; while other NP-h articles have been cited for less than or equal to h times, therefore, the scientist's H-index is h, which is demonstrated by the following formula:

$$[\max\{c \mid n(x) \geq c, d(y) \cdot S(y) \geq c\}] .$$

Among which $n(x)$ is the number of neighbor nodes of x, y is any neighbor node of x, $d(y)$ is the node degree of y, $s(y)$ is the strength connecting the edges of xy, and [z] is the maximum integer that is not bigger than z.H-index measured the key parts of a dataset by a relatively natural method [2].

Based on the H-index, this paper proposed an improved cooperation index (hereinafter referred to as improved index) to measure the core strength of nodes and reflect the capacity of the nodes in the weighted undirected networks. It concentrated reflects the nodes' degrees, strength and the cooperative ability of neighbor nodes. We defined the index of nodes in the weighted undirected networks as that the strength of the nodes' edges and the nodes should be no less than the maximum integer of the nodes strength product of the neighbor nodes of c (defined the node strength as the sum of the strength of the nodes' edges). In one weighted cooperative network, if one scholar owns many cooperators and with high frequency, or these cooperators own strong cooperative ability (the strong cooperative ability here refers to the total frequency between the scholar and others, i.e. the nodes' strength in the network), so he/she owns high index [3]. H degree only considered the nodes' degrees and the edges' strengths, while neglected the influence of the neighbor nodes. However, the importance degree of the neighbor nodes is a significant index influencing the core of the nodes, which is the core of citation ranking and webpage ranking.

2 Related work

The improved index measured the nodes' cooperative ability in the weighted networks, considered the number of the cooperators and the cooperation frequency, as well as the cooperators' cooperative ability. The index is similar to H-index, while it effectively measured the cores of the above information-the number of the cooperators, the cooperative ability and the cooperation frequency, as well as balanced the sources of all kinds of information. Its ability to measure the cooperative ability in the weighted networks can not be replaced by other indexes [4, 5]. It can be found from the comparison with other famous literature metrologies that it is obviously more effective than the core of degree (the number of cooperators), the nodes' strength (the sum of the cooperation frequency), h degree (without cooperators' importance degree and I-index (without cooperation frequency) and wl-index (calculate the strength of neighbor nodes only by cooperation frequency). However, due to the different information used for calculating, it is impossible to directly compare the closeness centrality with between's and eigenvectors concentration. In the following parts, we analyzed the relevance between key indexes, reflected that these three indexes are of low relevance and proved the uniqueness of index.

Based on the factors influencing the cooperative ability, this paper comprehensively summarized the characteristics of H-index, and proposed improved cooperation index, which measured the nodes' cooperative ability in the weighted networks as well as considered the number of cooperators and the cooperation frequency, as well as the cooperative ability of the cooperators themselves.

Definition 1 (c-index): the cooperation index *c*(*x*) of the node *x* is the maximum integer c, the node *x* owns *c* neighbor node meets at least, sum of each node's strength and the node *x*'s strength connecting edges is not less than *c*.

The following parts demonstrated how to calculate index. Figure 1 gave an instance for a weighted network. The width of the connecting lines referred to the strength, labeling with figures. Figure 1 is a cooperator's network, and we calculated the index of scholar C. Scholar C owns four cooperators: A, B, D and E, who had been cooperated with C for 1, 2, 4 and 4 times. The node strength of A, B, D and E is 1, 2, 7(=4+2+1) and 11(=4+2+5). Then calculate the arithmetic product of each node's strength and the cooperation frequency, and ranking them in descending order (refer to Figure 2). It is obvious that C owns 3 cooperators at least, thus the index of scholar c is 3.

The greater the strength of the neighbor nodes, the greater the ability to communicate and the influence they will have. The index has taken the two into consideration and multiplies the two. Thus the index adopts H-index to balance the arithmetic product and nodes' degrees. It follows that the index is applicable to describe the effective communication ability between nodes, which is consistent with the cooperative ability.

In order to discuss the distribution of index, the following lemmas shall be referred to:

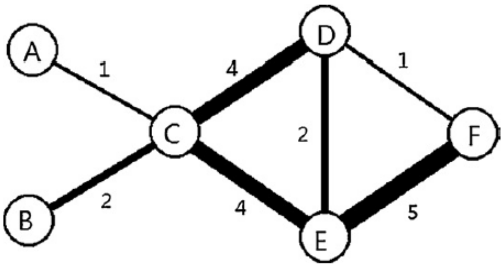
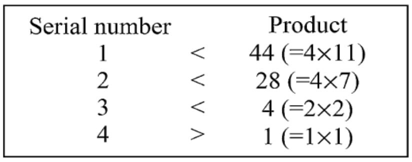


FIGURE 1 Weighted artificial network for analyzing index



Serial number		Product
1	<	44 (=4×11)
2	<	28 (=4×7)
3	<	4 (=2×2)
4	>	1 (=1×1)

FIGURE 2 The index calculation process for node C

3 The example and calculation

Lemma 1. If the probability distribution function *f*(*x*) of random variable X meets $f(x)ak^{-(y-1)}$, thus the $P(X \geq k)ak^{-(y-1)}$, which is also applicable to other cases [6].

Lemma 1 (distribution of index h) assumes one *a* - fat-tailed distribution: $P(n(x) \geq n) \approx\sim bn^{-a}$, citation score owns one fat-tailed $P(cit(y) \geq c) \approx\sim bc^{-\beta}$, if all publishing scores are independent from citation scores, thus $P(h(x) \geq k \cong k^{-a(\beta+1)}$.

Lemma 2. If the independent identically distributed variables X and Y are positive integers, and

$P(X \geq k) \approx\sim bk^{-a}$, thus $P(X+Y \geq k) \approx bk^{-a}$.

$$P(X+Y=K)=\sum_{X=1}^{k-1} P(X+Y=K, X=x)=$$

$$\sum_{X=1}^{k-1} P(X+Y=k \mid X=x)(\mathrm{X}=x)=$$

$$\sum_{x=1}^{k-1} P(Y=k-x)P(X=x) \approx$$

$$\mathrm{b}\sum_{x=1}^{k-1}(k-x^{-(a+1)}x^{-a+1} \approx ck^{-(a+1)})$$

Demonstration: According to the lemma 1, $P(X=k) \approx bk^{-(a+1)}$; and according to the total probability formula:

$$P(X+Y=K)=\sum_{X=1}^{k-1} P(X+Y=K, X=x)=$$

$$\sum_{X=1}^{k-1} P(X+Y=k \mid X=x)(\mathrm{X}=x)=$$

$$\sum_{x=1}^{k-1} P(Y=k-x)P(X=x) \approx$$

$$\mathrm{b}\sum_{x=1}^{k-1}(k-x^{-(a+1)}x^{-a+1} \approx ck^{-(a+1)})$$

According to the lemma 1, $P(X+Y \geq k) \approx bk^{-a}$.

The following is the distribution of index:

Lemma 3: (distribution of index) in the weighted undirected network G, all edges' strengths are mutually independent, the weight distribution of some S edges are submitted to $P(S \geq k) \approx bk^{-\beta}$, and all the nodes' degrees are independent. The distribution of some nodes' degrees meets $P(d(x) \geq k) \approx bk^{-a}$. When all edges' strengths are independent from the nodes' degrees, $P(c(x) \geq k) \cong k^{-a(\beta+3/2)}$ [7].

Demonstration: firstly, calculate the distribution of nodes' degrees:

Assume the number of the network G's nodes as N; S_x and *d*(*x*) are the node's strength and node's degree respectively. If *n*>*k*, then $P(S_x = k, d(x) = n) = 0$, thus it can be calculated according to the total probability formula and lemma 2 that:

$$P(S_x = k) = \sum_{n=1}^{\min(N-1,k)} P(S_x = k \mid d(x) = n) \cdot P(d(x) = n) \approx$$
$$\approx b \sum_{n=1}^{\min(N-1,k)} k^{-(\beta+1)} n^{-a} = bk^{-(\beta+1)} \sum_{n=1}^{\min(N-1,k)} n^{-a} \approx bk^{-(\beta+1)}$$

Therefore $P(S_x - S \geq k) \approx bk^{-\beta}$.

Assume that S is the edges' strengths of some nodes, and Sx is the strength of node x. now we'll calculate the distribution of SSx. Apparently, *S*_x>=*S*. According to the demonstration of (1), we can obtain that $P(S_x - S \geq k) \approx bk^{-\beta}$,

$$P(S_x S \geq k) = \sum_{y=1}^{+\infty} P(S_x S \geq k, S = y) = \sum_{y=1}^{+\infty} P(S_x \geq \frac{k}{y} \mid S = y \cdot$$
$$P(S = y) = \sum_{y=1}^{+\infty} P(S_x - S) \geq \frac{K}{y} - y) P(S = y) =$$
$$\sum_{y=1}^{[\sqrt{k}]-1}]\frac{k}{y} - y[^{-\beta} y^{-(\beta+1)} + \sum_{[\sqrt{k}]}^{+\infty} y^{-(\beta+1)} =$$
$$\sum_{y=1}^{\sqrt{k=1}} (\frac{k}{y} = y)^{-\beta} y^{-(\beta+1)} + \sum_{\sqrt{k}}^{+\infty} y^{-(\beta+1)} \approx$$
$$b_1 k^{-(\beta+1)/2} + b_2 k^{-\beta} \approx bk^{-(\beta+1)/2}$$

The operator [*x*] is the maximum integer that is no bigger than *x*,]*x*[is the minimum integer that is no smaller than *x*, and "*x*=" refers to approximately equal.

(1) according to (2), $P(S_x - S \geq k) \approx bk^{(-\beta+1)/2}$ is established; according to $P(d(x) \geq\sim k) \approx bk^{-a}$ and lemma 1, $P(c(x) \geq k) \cong k^{-a(-\beta+3)/2}$ is established.

Lemma 2 demonstrated that within the weighted free networks (the edges' strengths are natural numbers), if all edges are mutually independent and meet the power law of *β*, all nodes' degrees are mutually independent and meet the power-law distribution of *β*, and all nodes' degrees and strengths are mutually independent, thus the index meets the power law of $\alpha(\beta+3)/2$ [8].

Therefore, we analyzed the basic characteristics of the cooperator's network and studied the node degree (*d*), edge strength and node strength (*S*). After the LOG-LOG conversion of all indexes' cumulative probability distribution, we found that these indexes' distributions are approximately submitted to the power law. Therefore, cooperator's network can be taken as a strengthened scale-free network [9, 10].

Then we calculated all the nodes' index (*c*) in the cooperation network, and then had double logarithmic transformation on the cumulative probability distribution of *c*-index. Here the cumulative probability distribution of the index can be defined as follows:

$$F(c) = \Pr\{c - index \geq c\} = \frac{\#\{x : c(x) \geq \text{c}\}}{b}$$

"#" refers to the number of elements in the set. Other indexes of the cumulative probability distribution are defined similarly [11-13]. The distribution of *c*-indexes after transformed data regression approximates power-law distribution. All models and coefficients passed the tests, which is as shown in Figure 3.

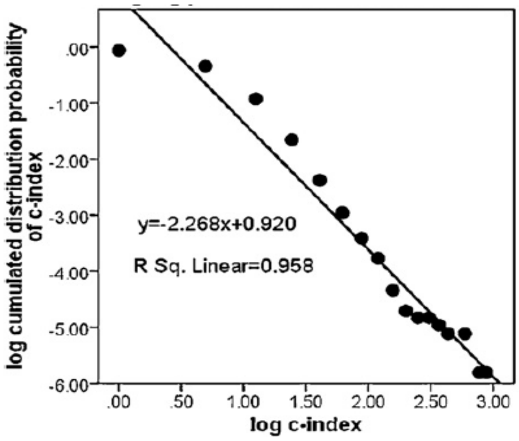


FIGURE 3 Distribution of the index

Figure 4 demonstrated the cooperator's subnet, including 50 authors of the maximum indexes. Their indexes are greater than or equal to 12. In the figure, the wider the edge is, the stronger the strength of the connecting node is indicated; and the larger the node is, the higher the index is indicated. The index value is marked on the node, and the strength value of the edge is marked on the edge as well, i.e. cooperation duration. It follows that even if a part of the network is able to demonstrate the authors of high index, more cooperators and stronger cooperation strength and strong cooperation status.

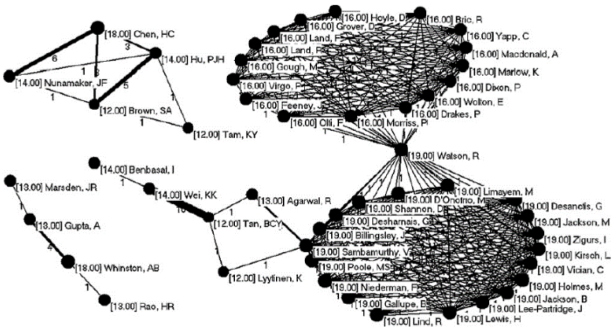


FIGURE 4 50 seculars according to the improved index

4 Analysis of improved index based compared to the other index

To compute the nodes based on H - type factor: l - the index (l), wl - index (wl) and H - degree (d_h). The comprehensive inspection correlations of the relationship between these factors were calculated. Because of these factors is grade data, the Spearman rank correlation was used, the factors of the results are shown in Table 1.

In the Table 1, because of the degree of relevance between node strength reached 0.981, so the weighted network and the corresponding non-weighted network.

TABLE 1 The analysis of Spearman correlation coefficient between several indexes

	c	dh	wl	l	ev	bw	cl	s	d
c	-	0.485**	0.999**	0.996**	0.426**	0.453**	0.601**	0.933**	0.955**
Other h-index									
dh		-	0.482**	0.469**	0.299**	0.375**	0.450**	0.554**	0.502**
wl			-	0.996**	0.426**	0.450**	0.599**	0.931**	0.954**
l				-	0.420**	0.432**	0.594**	0.924**	0.948**
The classic centricity metrics									
ev					-	0.399**	0.817**	0.451**	0.439**
bw						-	0.476**	0.629**	0.624**
cl							-	0.629**	0.620**
s								-	0.981**
d									-

Difference is not big. H - types except h - degree indexes (l - index, wl - index, c - index and strongly correlated to the node degree, node strength. Since the edge strength for natural Numbers, network of average intensity is close to the minimum while strength 1 value of 1, the index of many nodes (such as the node of isolated nodes and the degrees of 1) is same. If the network relationship of M scholars were consisted only through one cooperative partnership, then their c - index, wl - index, l - index and the node degrees are equal. Many analysis show that most scholars cooperated only once, so many nodes in accordance with the following situations, c=l=wl=d. Strong related node degree and strength of the node is c - index, l - index, wl - index, node degree, node strength is the main reason for the high correlation. L - the index with the three classic center index (close to the center (cl), middle center (bw), characteristic vector center degrees (ev) is relatively weak correlation, and strongly correlated to the degree, this has to do with Korn and others [14] of the results. C - Index, IC - index and l - index, wl - index has a maximum correlation. For example, c - index, wl - the index is 0.999, which is caused by the characteristics of the network edge right. Because the weighted network with no significant differences is corresponding to unweighted network, so the proposition 3 shows that many nodes in the network will meet c=wl=l. In addition, obviously, if the node of edge strength are all 1, then this node c=wl. Besides, there are many isolated nodes, in the network, the node degrees =1, so c=IC=l=wl. All these lead to the high correlation.

5 Conclusions

This paper put forward an improved factor for the centers of classic metric (center degrees, proximity, mediation, characteristic vector center) and other factor h - type (l - index, wl - index and h - degree). The index of comprehensive use the number of neighbor nodes, adjacency node connection strength and the concentration to measure the concentration, which provide more accurate information to the cooperate ability [15].

In the scale-free weighted networks, the index follows a power law, in the weighted network the cooperation ability of nodes is up to the number of neighbor nodes (nodes), cooperation intensity (intensity) and the importance of cooperate or collaborator (adjacency node connection), ability to cooperate for the three monotonous reduction function.

The improved index is used to simply node cooperation measurement. The definition of c-index should be defined as the product of strength and the adjacent point strength of h - index. Although the discussion of the type is suitable for cooperation network, the cooperation factor also could be used for other nodes in complex networks.

References

[1] Litvak N, Scheinhardt W R W, Volkovich Y V 2011 Probabilistic relation between In-Degree and PageRank *In Fourth International Workshop WAW 2006 Nov 30–Dec 1 2006, Banff Canada* 72–83

[2] Liu L G, Xuan Z G, Dang Z Y, Guo Q, Wang Z T 2011 Weighted network properties of Chinese nature science basic research *Physica A-Statistical Mechanics and Its Applications* **377**(1) 302-14

[3] Ma N, Guan J, Zhao Y 2008 Bringing PageRank to the citation analysis *Information Processing and Management* 44 800-10

[4] Milgram S 2013 The small world problem *Psychology Today* 2 60-7

[5] Nascimento M A, Sander J, Pound J 2003 Analysis of SIGMOD's coauthorship graph *SIGMOD Record* **32**(3) 8-10

[6] Newman M E J 2001 Scientific collaboration networks: I. Network construction and fundamental results *Physical Review E* 64 016131

[7] Newman M E J 2013 The structure of scientific collaboration networks *Proceedings of the National Academy of Science of the United States of America* **98**(2) 404-9

[8] Page L, Brin S 1998 The anatomy of a large-scale hypertextual Web search engine *Computer Networks and ISDN Systems* 30 107-17

[9] Paullay I M, Alliger G M, Stoneromero E F 2013 Constructvalidation of 2 instruments designed to measure job involvement and work centrality *Journal of Applied Psychology* **79**(2) 224-8

[10] Perra N, Fortunato S 2008 Spectral centrality measures in complex networks *Physical Review E* 78 036107

[11] Pinski G, Narin F 1976 Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics *Information Processing and Management* 12 297-312

[12] Price J D S 1965 Networks of scientific papers *Science* 149 510-5

[13] Newman M E J 2010 The structure and function of complex networks *SIAM Review* **45**(2) 167-256

[14] Newman M E J 2005 A measure of betweenness centrality based on random walks *Social Networks* 27 39-54

[15] Nisonger T E, Davis C H 2012 The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl–Davis study *College & Research Libraries* 66 341-77

Authors

	Dai Lu, born in 1988, P.R. China. **Current position, grades**: currently working at DongGuan Univercity of Technology. **University studies**: PhD degree from Wuhan University, in 2013. **Scientific interest**: cloud computing, evolutionary algorithm, social network and text mining.
	Li Guangming, born in 1968, P.R. China. **Current position, grades**: vice professor at Dongguan University of Technology. **University studies**: BsC form Henan Normal University in 1995, ME degree form Zhengzhou University of Technology in 2000. **Scientific interest**: nonlinear circuit theory, chaotic signal processing, embedded systems design, social network and their applications.