Malaysian Journal of Library & Information Science, Vol. 16, no. 3, December 2011: 9-16

An empirical relation between k-shells and the h-index in scale-free networks

Fred Y. Ye[1,2] , Star X. Zhao[1] and **Ronald Rousseau[3,4,5]**
[1]Zhejiang University, Department of Information Resource Management, Hangzhou, CHINA
[2]Institute of Scientific and Technical Information of China, Beijing, CHINA
[3] KHBO (Association K.U.Leuven), Faculty of Engineering Technology,
Zeedijk 101, B-8400 Oostende, BELGIUM
[4] K.U.Leuven, Dept. Mathematics,
Celestijnenlaan 200B, B-3001 Leuven (Heverlee), BELGIUM
[5] Antwerp University (UA), IOIW, Venusstraat 35, B 2000 Antwerp, BELGIUM
e-mail: ronald.rousseau@khbo.be

ABSTRACT

After clarifying the definitions of h-index and k-shells in a graph, it is shown that the largest k value for which there exists a non-empty k-shell, denoted as $k_{max}(G)$, satisfies the relation $k_{max}(G) \leq h(G)$, where h(G) is the degree h-index of graph G. Next we determine an empirical relation between the h-index, the number of nodes in a small scale-free network, i.e. with maximum degree centrality < 100, and the coreness and degree centrality of its nodes. In this contribution we embed the information sciences among other fields involved in network studies.

Keywords: h-index; k-shells; Power laws; Graphs; Scale-free networks

INTRODUCTION

Since its introduction in 2005 the *h*-index or Hirsch index (Hirsch 2005) has been the topic of several hundreds of articles, often, but not always, within a publication-citation context. As citations can be considered links in a network of articles it is not surprising that the notion of an *h*-index has been extended to a network context (Schubert et al. 2009). The *h*-index has been introduced as an indicator to evaluate the lifetime achievements of a scientist (Hirsch 2005). In that context it is defined as the largest natural number *h* such that the articles ranked 1 to *h* have at least *h* citations. This definition has been generalized, modified and adapted to many other cases besides scientists. For more information on the *h*-index, its uses and generalizations we refer the reader to the review articles by Alonso et al. (2009) and by Egghe (2010).

Recently also the notion of an *h*-core has been extended to a network context (Rousseau and Ye 2011). Yet, in networks another type of core, namely a *k*-core has been introduced many years ago (Seidman 1983; Bollobás 1984). In this article we put forward an empirical relation between the *h*-index of a network and the coreness of its nodes in small (maximum degree centrality smaller than 100) scale-free networks. Throughout this article the terms network and graph are used as synonyms.

K-CORE AND *K*-SHELL OF A GRAPH

A *k*-core of a graph G was defined by Seidman (1983), Bollobás (1984) as well as Wasserman and Faust (1994) as a maximal connected subgraph in which each node has at least degree *k*. We will ignore the connectedness requirement, leading to a unique *k*-core. A *k*-shell in the sense of Carmi et al. (2007) can be constructed as follows: a) First, remove from a graph G all nodes of degree less than *k*. Remove also their links. In this way some of the remaining nodes may now have less than *k* links; (b) Then remove these nodes too, and so on until no further removal is possible. The result, if it is non-empty, is the *k*-shell, which is often also referred to as *k*-core (Yin et al. 2006). A *1*-shell is just a component (Hanneman and Riddle 2005). The largest *k* value for which there exists a non-empty *k*-shell is then denoted as $k_{max}(G)$. The corresponding subgraph of a graph G is called the k_{max}-shell or the nucleus. If this nucleus is connected then it coincides with the notion of a *k*-core as defined by Seidman. Carmi et al. (2007) define a *k*-crust as the union of all shells with indices larger than or equal to *k*. In this article the term *h*-shell denotes a *k*-shell where $k = h(G)$, the *h*-index of the graph G. Because of its construction Dorogovtsev et al. (2006) refer to the result of the procedure leading to *1*-, *2*- , … *k*-shells in a given graph as a Russian nesting doll and the action itself can be compared with peeling an onion (Wuchty and Almaas 2005).

Clearly, following the definition used by Carmi et al. (2007), a *k*-shell may consist of different disconnected components, such that each component contains at least *k+1* nodes. If a graph has a *k*-shell then it also has an *m*-shell for $m = 1, \ldots, k$ as each *k*-shell is a subgraph of each *m*-shell ($m < k$). These *m*-shells may, however, coincide, as is the case for a complete graph. A node *u* has *coreness c* if it belongs to the *c*-shell but not to a $(c + 1)$-shell.

k-cores and *k*-shells are useful in mathematics as well as the natural and social sciences, particularly in graph theory (Pittel et al.1996), network visualization (Alvarez-Hamelin et al. 2006), social networks (Moody 2001) and protein analysis (Bader and Hogue 2002). Carmi et al. (2007) used it to study the Internet topology, while Leydesdorff and Wagner (2008) studied international science collaboration. They found that in the network of international scientific collaborations (with countries as nodes) the size of the nucleus grew from 35 (1995) to 53 (2000) to 64 (2005) countries. Biology, in particular protein analysis, was one of the first fields where the *k*-shell decomposition was applied. Besides the Bader and Hogue article (2002), we also mention the prediction of the protein functions by Altaf-Ul-Amin et al. (2003). Another interesting work in the same area is Wuchty and Almaas' (2005) study on the centrality of the protein function.

SIMPLE THEORETICAL RELATIONS BETWEEN THE *H*-INDEX AND *K*-SHELLS

In this short section we mention some simple theoretical relations between $h(G)$ and $k_{max}(G)$. The most important one is given in the next proposition.

Proposition 1*:* The *h*-index of a *k*-shell in a graph G is larger than or equal to *k*. Consequently $k_{max}(G) \leq h(G)$.

Proof : As a *k*-shell has at least *k+1* nodes, each with degree at least *k*, the *h*-index of this *k*-shell is at least *k*.

It is possible that this *h*-index is exactly *k*, as shown by a triangle, which is a 2-shell and its *h*-index is equal to 2. However, this equality is not always true. In Figure 1, we show a graph G for which $k_{max}(G)$ *is* 2 while *h(G)* is 3. Larger differences are even possible. Indeed, for a tree $k_{max}(G)$ equals 1. If this tree is an *n*-ary tree (i.e. each node, except the terminal ones, has *n* children) and has depth at least 2, its *h*-index is *n*.

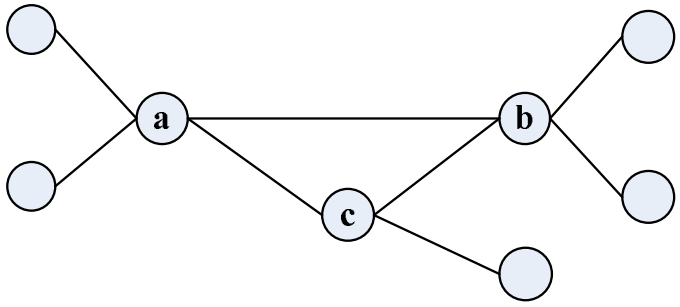

Figure 1: Graph G with h(G) = 3 and $k_{max}(G)$=2

Proposition 2 *:* If G is a complete graph of order *n*, then it is its own nucleus, with $k_{max}(G) = n-1$. The coreness of each node is *n-1*, and the *h*-index of this complete graph is also *n-1*.

Proof: this follows immediately from the definitions.

The next section contains an empirical result relating h-index to coreness and degree centrality in power-law dominated systems or scale-free networks, which will reveal that there is relation between traditional network parameters and the *h*-index.

EMPIRICAL RELATIONS IN SCALE-FREE NETWORKS

According to practical Internet data (Mahadevan et al. 2005) coreness (*c*) is linked to degree (*d*) by a power law, at least for networks with nodes with a small degree (maximum degree centrality *d < 100*). So, we can write the following formula:

$$c(d) = d^{\beta} \tag{1}$$

where, according to (Mahadevan et al. 2005) the power-law exponent *β* can be estimated between 0.6 and 1.1. For large *d* (much larger than 100) *c* seems to be constant: *c = C*. (We note that the discussion of coreness was eliminated from the published version of Mahadevan et al. (2006)).

Consider a graph and let *f(d)* denote the number of nodes of degree *d*, then the node degree distribution, denoted as *P(d),* is the probability that a node has degree *d*: *P(d) = f(d)/N*, where *N* denotes the total number of nodes in the graph. Rousseau (1997), on a small sample, and Faloutsos et al. (1999), on a large scale, have shown that the Internet's degree distribution follows a power law. Generally, the degree distribution of nodes in a scale-free network obeys a power-law (Barabási and Albert 1999; Newman 2003). Then the number of nodes of degree *d* is given by

$$f(d) = ad^{-\gamma} \tag{2}$$

where *a, γ > 0* are constants. In classical informetrics, equation (2) is known as Lotka's law.

Egghe and Rousseau (2006) derived that the *h*-index in power-law dominated informetrics can be written as

$$h = N^{1/\alpha} \tag{3}$$

in which α > 1, is the power-law coefficient for the degree distribution and *N* is the total number of nodes excluding singleton nodes, i.e. nodes with degree zero. Setting *α=γ* and combining formulas (1), (2) and (3), we obtain

$$ad^{-\alpha} = ac(d)^{-\alpha/\beta} = ac(d)^{-(\ln N/\ln h)/\beta} \tag{4}$$

Taking logarithms in equation (4) yields

$$\ln(h) = \frac{\ln(c(d)\ln(N)}{\alpha\beta\ln(d)} \tag{5}$$

or

$$\ln(c(d)) = \frac{\alpha\beta\ln(h)\ln(d)}{\ln(N)} \tag{6}$$

Setting *m= 1/(α.β)* we find

$$\ln(h) = m\ln(N)\frac{\ln(c(d))}{\ln(d)} \tag{7}$$

Note that m is a constant for a given network and that the base of the logarithm does not matter in these equations. Equations (5), (6), (7) are empirical nonlinear relations between a network's *h*-index and its nodes' coreness *c(d)= k_{max}(G)* in a scale-free network with node degree at most 100.

While there are a lot of nodes with different coreness and degree, each network has its unique *h*-index. When *N* is fixed, then *ln(h)/ln(N)* will be a constant. So, in such a scale-free network, we have the relation

$$\frac{m\ln(c(d))}{\ln(d)} = \frac{\ln(h)}{\ln(N)} \tag{8}$$

A TEST IN A REAL-WORLD FRAMEWORK

There exist many articles modelling networks as scale-free networks, involving power law relations (e.g. Egghe 2008) without giving concrete examples. Moreover, it is well-known that the power law relation, if it holds, only holds approximately, or only in the tail. The power law relation for coreness, in particular, has only been established for three networks. So, although equation (8) is correct if all assumptions underlying it hold, we would like to

check if it holds, at least approximately, for a real network. For this purpose we collected keywords from five prominent mathematical journals (*Annals of Mathematics*, *Communications on Pure and Applied Mathematics*, *American Journal of Mathematics*, *Pacific Journal of Mathematics*, *Quarterly Journal of Mathematics*) and constructed a co-keyword network, focussing on the top 3% keywords (the most occurring ones). This yielded 60 nodes, shown in Figure 2.

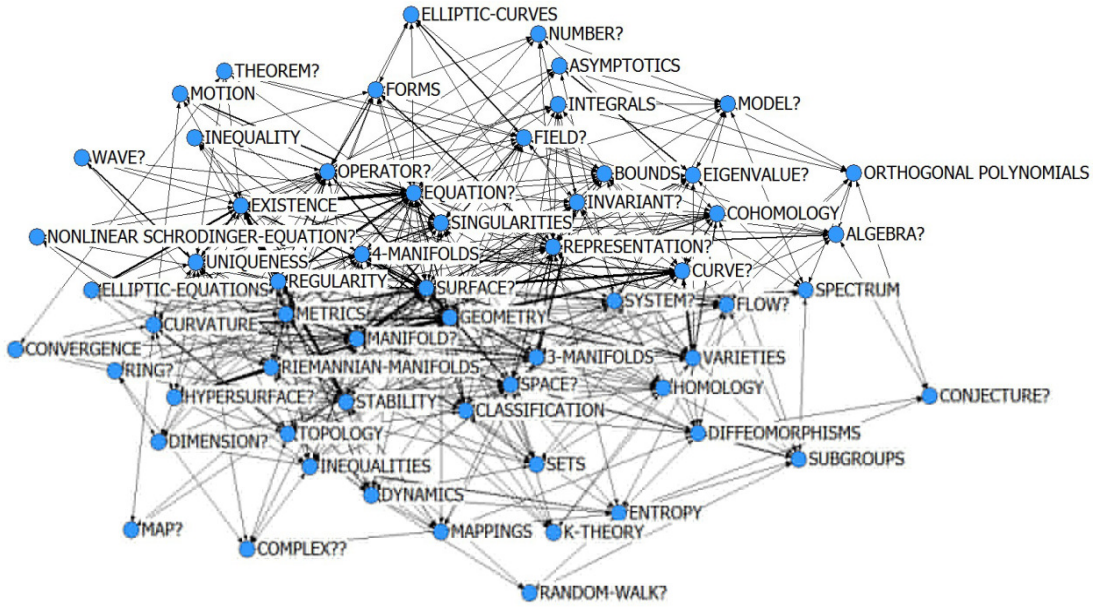

Figure 2: A Co-keyword Network as an Example of a Small Scale-free Network

The *h*-index of this network is 23 and $ln(h)/ln(N) = ln(23)/ln(60)$ = 0.7658. For each node we know its degree and its coreness. Hence it would theoretically be possible to determine the constant *m*, using equation (8). Figure 3 shows that the resulting value for *m* is indeed approximately constant as the value for *m* calculated for each node lies in a strip between 0.8 and 1.2. This seems to confirm that our model is an acceptable approximation. However, we looked somewhat deeper in this and found that if we rank nodes according to their degree centrality (Figure 4), there seems to be a decreasing relation between degree centrality and the obtained *m*-value. This is an interesting finding whose solution we have to leave as an open problem.

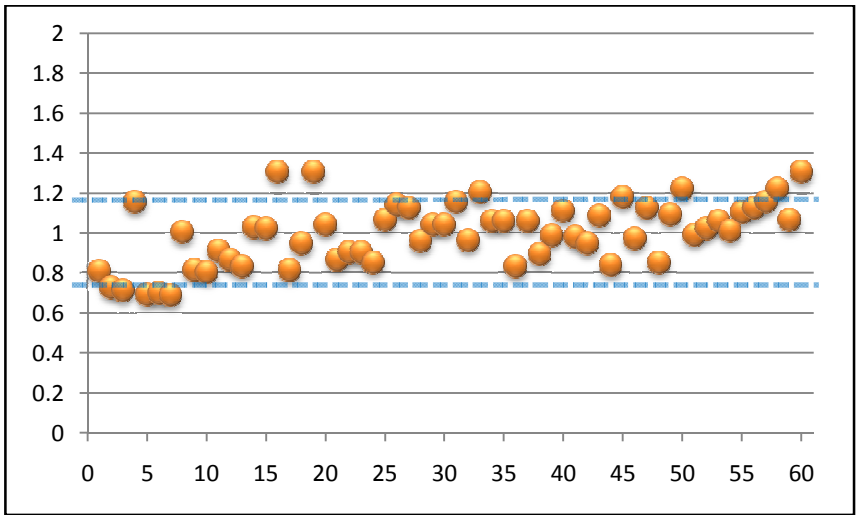

Figure 3: Values for *m* Derived from Different Nodes

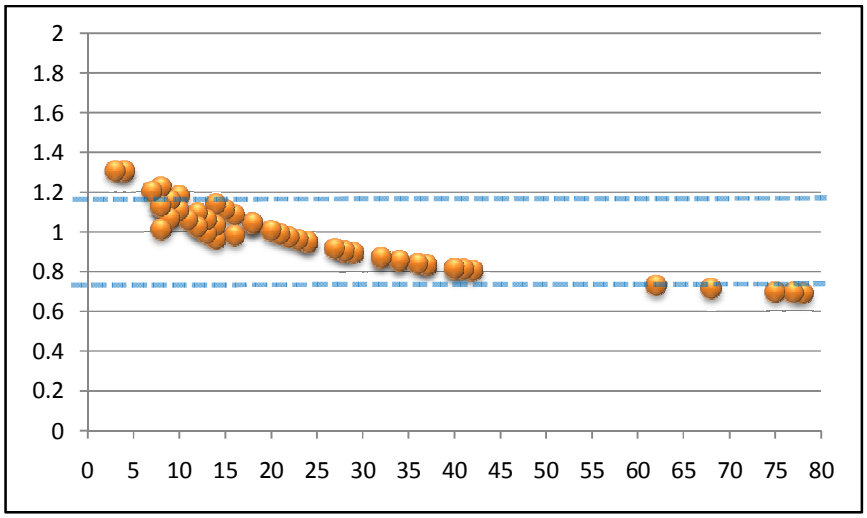

Figure 4: Obtained *m* Values as a Function of Node Degree

CONCLUSION

We have shown that the largest *k* value for which there exists a non-empty *k*-shell, denoted as $k_{max}(G)$, satisfies the relation $k_{max}(G) \leq h(G)$, where $h(G)$ is the degree *h*-index of graph *G*. Next we recalled the notion of coreness of a node in a network and determined an empirical relation between the *h*-index, the number of nodes in a scale-free graph, i.e. with maximum degree centrality < 100, and the coreness and degree centrality of its nodes. Studying coreness and in particular its relation with the *h*-index in scale-free networks opens a new line of inquiry in the information sciences. It is another way of giving the information sciences its rightful place among other fields involved in network studies. We are convinced that our study will stimulate further ones involving node coreness.

ACKNOWLEDGEMENT

The authors acknowledge NSFC Grants (No. 70773101 and No. 7101017006) for supporting their cooperation. Fred Y. Ye thanks for the financial support from Humboldt University for staying in iFQ for collaborative studies, which moreover provided him the opportunity to visit Ronald Rousseau (KHBO) in Belgium. The authors thank Mr. Paul L. Zhang for his programming work in relation to the keyword network studied in this contribution and Mr. Raf Guns (UA) for pointing out an error in an earlier version.

REFERENCES

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E. and Herrera, F. 2009. H-index: a review focused in its variants, computation and standardization for different scientific fields, *Journal of Informetrics*, Vol. 3, no 4: 273-289.

Altaf-Ul-Amin, MD., Nishikata, K., Koma, T., Miyasato, T., Shinbo, Y., Arifuzzaman, Md., Wada, C., Maeda, M., Oshima, T., Mori, H., and Kanaya, S. 2003. In M. Gribskov, M. Kanehisa, S. Miyano and T. Takagi (eds.).Prediction of protein functions based on K-cores of Protein-Protein interaction networks and amino acid sequences. *Genome Informatics* 14: 498-499. Tokyo: Universal Academy Press.

Alvarez-Hamelin, J.I., Dall'Asta, L., Barrat, A., and Vespignani, A. 2006. Large scale networks fingerprinting and visualization using the *k*-core decomposition. In Y. Weiss, B. Schölkopf, and J. Platt (eds.). *Advances in Neural Information Processing Systems* 18*:* 41-50, Cambridge (MA): MIT Press.

Bader, G.D. and Hogue, C.W.V. 2002. Analyzing yeast protein–protein interaction data obtained from different sources, *Nature Biotechnology*, Vol. 20 : 991-997.

Bollobás, B. 1984. The evolution of sparse graphs. In B. Bollobás (ed.). *Graph Theory and Combinatorics: Proc. Cambridge Combinatorial Conf. in honour of Paul Erdös* Academic Press, NY: 35-57.

Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. and Shir, E. 2007. A model of Internet topology using k-shell decomposition, *Proceedings of the National Academy of Sciences USA*, Vol. 104, no 27: 11150-11154.

Dorogovtsev, S.N., Goltsev, A.V. and Mendes, J.F.F. 2006. k-Core organization of complex networks, *Physical Review Letters*, Vol. 96, 040601. Doi: 10.1103/PhysRevLett.96.040601.

Egghe, L. 2008. A model for the size-frequency function of coauthor pairs. *Journal of the American Society for Information Science and Technology*, Vol. 59, no 13: 2133-2137.

Egghe, L. 2010. The Hirsch-index and related impact measures, *Annual Review of Information Science and Technology*, Vol. 44 : 65-114.

Egghe, L. and Rousseau, R. 2006. An informetric model for the Hirsch-index, *Scientometrics*, Vol. 69, no 1: 121-129.

Hanneman, R. A. and Riddle, M. 2005. *Introduction to social network methods*. Riverside, CA: University of California, Riverside. Available at: http://faculty.ucr.edu/ ~hanneman/)

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the USA*, Vol. 102, no 46: 16569-16572.

Leydesdorff, L. and Wagner, C.S. 2008. International collaboration in science and the formation of a core group, *Journal of Informetrics*, Vol. 2, no 4: 317-325.

Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., Claffy K.C. and Vahdat, A. 2005. Lessons from three views of the Internet topology. arXiv: cs/0508033v1

Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., Claffy K.C. and Vahdat, A. 2006. The Internet AS-level topology: three data sources and one definitive metric, *ACM SIGCOMM Computer Communication Review*, Vol. 36, no 1: 17-26.

Moody, J. 2001. Peer influence groups: identifying dense clusters in large networks, *Social Networks*, Vol. 23, no 4: 261-283.

Pittel, B., Spencer, J. and Wormald, N. 1996. Sudden emergence of a giant k-core in a random graph, *Journal of Combinatorial Theory Series B*, Vol. 67, no 1: 111-151.

Rousseau, R. and Ye, F.Y. 2011. Subgraphs derived from the Hirsch core in undirected, unweighted networks, *ISSI Newsletter*, Vol. 7, no 1: 5-9.

Schubert, A. 2009. Using the h-index for assessing single publications. *Scientometrics*, Vol. 78, no 3: 559-565.

Schubert, A., Korn, A. and Telcs, A. 2009. Hirsch-type indices for characterizing networks, *Scientometrics*, Vol. 78, no 2: 375-382.

Seidman, S.B. 1983. Network structure and minimum degree, *Social Networks,* Vol. 5, no 3: 269-287.

Wasserman, S. and Faust, K. 1994. *Social network analysis*. Cambridge University Press.

Wuchty, S. and Almaas, E. 2005. Peeling the yeast protein network, *Proteomic*s, Vol. 5, no 2: 444-449.

Yin, L.C., Kretschmer, H., Hanneman, R.A. and Liu, Z.Y. 2006. Connection and stratification in research collaboration: an analysis of the COLLNET network, *Information Processing & Management*, Vol. 42, no 6: 1599-1613.