Chapter 13

Progress of Studies of Citations and PageRank

Wataru Souma and Mari Jibu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77389

Abstract

A number of citations have been used to measure the value of paper. However, recently, Google's PageRank is also extensively applied to quantify the worth of papers. In this chapter, we summarize the recent progress of studies on citations and PageRank. We also show our latest investigations of the citation network consisting of 34,666,719 articles and 591,321,826 citations. We propose the generalized beta distribution of the second kind to explain the distribution of citation and introduce the stochastic model with aging effect and super preferential attachment. Furthermore, we clarify the positive linear relation between citations and Google's PageRank. By using this relationship as the benchmark to classify papers, we extract extremely prestigious papers, popular papers, and rising papers.

Keywords: citation, PageRank, SCI-E, fat tail, stochastic model, prestigious papers, popular papers, rising papers

1. Introduction

Citation analysis has a long history. Recently, Hou [1] applied the new method called the reference publication year spectroscopy (RPYS) to 2543 papers including 56,392 references regarding citation analysis in Science Citation Index Expand (SCI-E) and Social Science Citation Index (SSCI) data from 1970 to July 2016. This investigation clarified that the development of citation analysis is divided into five periods: before 1990, 1901–1950, 1951–1970, 1971–2000, and 2001–2016. In this chapter, we focused on the distribution of citations which were introduced by Price [2] and extensively investigated in the third period, that is, 1950s–1970s. In this chapter, we consider that the number of citations expresses the popularity of papers.

The fifth period, that is, 2001–2016, is characterized by a period of rapid expansion and diversified directions. In this period, many conceptions have been introduced, for example, scientific

evaluation indices, citation networks, information visualization, and citing behaviors. A variety of new impact measures has been proposed based on social network analysis in sociology and of network science originated from physics, mathematics, and information science. Bollen [3] summarized 39 impact measures and investigated the correlation between them by using the principal component analysis. Then, Bollen [3] indicated that the notion of scientific impact is a multidimensional construct that cannot be adequately measured by any single indicator, although some measures are more suitable than others.

In this chapter, we focus on the Google's PageRank which is first proposed by Brin and Page [4] to obtain the list of useful web pages for queries by users. Thus, if we define the usefulness of web page as the number of links cited by the other web pages, the search engine should propose the list of portal sites, that is, popular web pages. Hence, this list is useless for web users. To overcome this problem, based on the concept of vote, Brin and Page [4] defined the usefulness of web pages as the number of votes from the linking web pages. In the algorithm of Google's PageRank, the number of ballets is proportional to the usefulness of the web page, that is, the useful web page has many ballets. As a result, the useful web page collects votes from the useful web pages. Thus, the Google's PageRank expresses the prestige of web pages. We consider that this characteristic of Google's PageRank is valid for the case of citation network.

This chapter is organized as follows. In Section 2, we explain characteristics of dataset used in this chapter. The distribution of citation and the stochastic model of citation network are elucidated in Section 3. In Section 4, we introduce Google's PageRank and calculate it. We consider the correlation between citation and PageRank in Section 5. Section 6 is devoted to conclusions.

2. Data

In this chapter, we use Science Citation Index Expand (SCI-E) provided by Clarivate Analytics Co., Ltd. This dataset contains bibliographic information of scientific papers published from 1900 to the present. However, due to limited research budget of authors, we use the dataset from 1981 to 2015 in this chapter. This dataset contains 34,666,719 papers and 591,321,826 citations.

In this chapter, we denote the number of papers published in the year t as $n(t)$. **Figure 1** depicts the change of $n(t)$. In this figure, $n(t)$ almost monotonically increased from 1981 to 2013 and decreased after 2013. However, this behavior of $n(t)$ is fake. This is because the dataset was made at the beginning of 2016 and it partially contains papers published in 2014 and 2015. It takes a few years for all the papers to be included in SCI-E.

If we consider papers as nodes and regard citations from a citing paper to a cited paper as directed links, we can consider the dataset of citations as a directed network. We call such a network as the citation network. The citation network consists of many connected components. We denote the number of nodes contained in connected components as c and represent a frequency of c as $F(c)$. **Figure 2** depicts $F(c)$. We can find that there is the largest connected

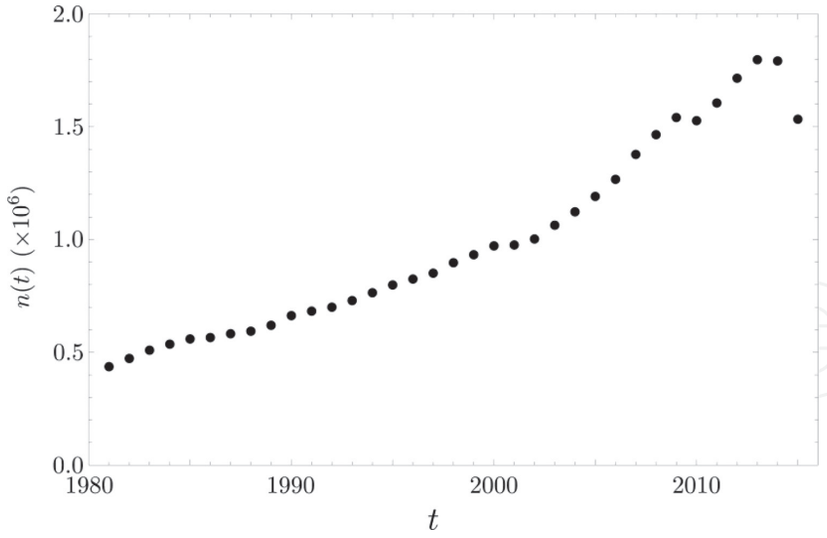

Figure 1. Yearly change of the number of e-articles.

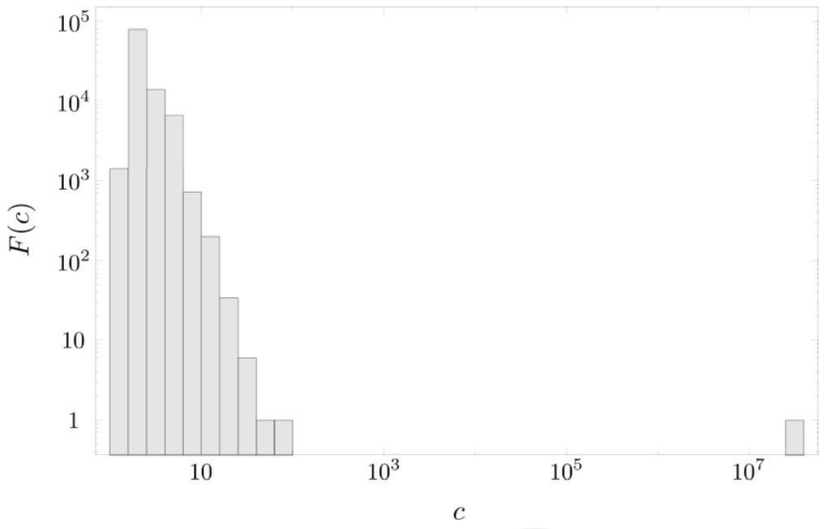

Figure 2. Distribution of the size of connected components.

component. This largest connected component consists of 34,428,322 nodes which are 99.3% of the total number of papers contained in the dataset, and of 591,177,607 links which are 99.98% of the total number of citations contained in the dataset. In the following section, we focus on the largest connected component.

3. Distribution and dynamics of citations

In this chapter, we argue for the distribution of the citations and stochastic models which lead to the citation network.

3.1. Distribution

The number of citations is represented by the number of in-degree, k, of the corresponding nodes. **Figure 3** is a double-logarithmic scale plot of the rank size distribution, $R(k)$, of citations. The right-tail part of the distribution decreases almost monotonically. This means that this part follows a power-law distribution, that is, $R(k) \propto k^{-\mu}$. Here, the exponent μ is called Pareto exponent originated in the name of Italian economist Vilfredo Pareto. The dashed line in **Figure 3** is the reference line which is the power law distribution with $\mu = 2$, that is, $R(\mu) \propto k^{-2}$.

Pareto [5] first investigated the fat-tail behavior of the right-tail part of personal income and wealth distributions. After Pareto, many types of distribution functions have been mainly proposed in the field of economics, especially in the investigation of personal income distribution (e.g., see [6, 7]). On the other hand, in the field of scientometrics, Price [2] first applied the power law distribution to the citation network and found that the distribution of the number of citing (the number of out-going degree in terms of network science) follows the power law distribution with $\mu = 1$ and that of the number of citations (the number of incoming degree in terms of network science) obeys the power law distribution with $\mu = 1.5$ or $\mu = 2$. The latter result is same as the reference line in **Figure 3**.

Rednar [8] investigated papers published in 1981 and cataloged by the Institute for Science Information (783,339 papers) and 20 years of publications in Physical Review D, vols. 11–50 (24,296 papers) and found that the right-tail part of both distributions of citation follows the power law distribution with $\mu = 2$. This result is same as Price [2] and the reference line in **Figure 3**. Rednar [9] investigated 110 years (from July 1893 through June 2003) of publications in Physical Review, the topical journals Physical Review A-E, Physical Review Letters, Review of Modern Physics, and Physical Review Special Topics: Accelerators and Beam (353,268 papers and 3,110,839 citations) and found that the entire distribution of the number of citation follows a log-normal distribution.

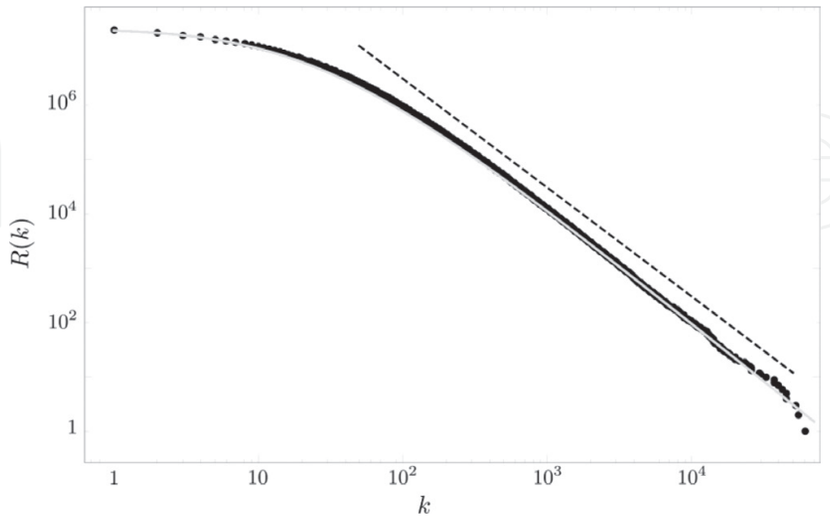

Figure 3. Rank size distribution, $R(k)$, of the number of citations, k.

Albarrán and Ruiz-Castillo [10] studied 5 years (1998–2002) of publications in Web of Science (3.7 million papers) and found that the power law distributions of the right-tail part of the distribution of citation are not rejected for 17 of the 22 scientific fields of Web of Science. Albarrán et al. [11] investigated same dataset of Albarrán and Ruiz-Castillo [10] and found that the power law distributions of the right-tail part of the distribution of citation are not rejected for 140 of the 219 scientific sub-fields of Web of Science. Recently, Brzezinski [12] investigated scientific papers published between 1998 and 2002 drawn from Scopus and found that the power law hypothesis is rejected for half of the Scopus field of science.

Although there are many researches besides the studies stated above, there are no studies that used vast amounts of data to approach the overall picture of citation distribution, like this chapter. The light gray line in **Figure 3** is the best fit by the generalized Beta distribution of the second kind (GB2) (or called the beta prime distribution) (e.g., see [13, 14]) with the probability density function:

$$f(k; a, b, \mu, \nu) = \frac{ak^{a\mu-1}}{b^{a\mu}B(\mu, \nu)}\left[1 + \left(\frac{k}{b}\right)^{a}\right]^{-(\mu+\nu)}, \tag{1}$$

with $a = 0.7$, $b = 15.2$, $\mu = 2.0$, $\nu = 3.0$. Here, $B(\mu, \nu)$ is the Beta function.

Table 1 depicts the top 20 papers of citation. In this table, r_k is the rank of citation, k is the number of citations at the beginning of 2016, and k', which is enclosed in parentheses, is the number of citations at the beginning of January 2018. The characteristics of this list are that the subjects of papers are almost Biochemistry & Molecular Biology and that the publication years of papers are relatively old.

r_k	$k(k')$	First author	Title	Journal, Year	Subject
1	60,967 (62,404)	P. Chomczynski	Single-step method of RNA isolation by …	Analytical Biochemistry, 1987	Biochemistry & Molecular Biology; Chemistry
2	55,143 (65,452)	A.D. Becke	Density-functional thermochemistry. 3…	Journal of Chemical Physics, 1993	Chemistry; Physics
3	52,035 (61,637)	C.T. Leer	Development of the Colle-Salvetti correlation…	Physical Review B, 1988	Physics
4	45,349 (64,127)	G.M. Sheldrick	A short history of SHELX	Acta Crystallographica Section A, 2008	Chemistry; Crystallography
5	44,915 (64,682)	J.P. Perdew	Generalized gradient approximation…	Physical Review Letters, 1996	Physics
6	42,407 (46,286)	J.D. Thompson	Clustal-W – Improving the sensitivity of …	Nucleic Acids Research, 1994	Biochemistry & Molecular Biology

r_k	$k(k')$	First author	Title	Journal, Year	Subject
7	39,281 (44,765)	S.F. Altschul	Gapped BLAST and PSI-BLAST: a new…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
8	37,133 (48,832)	S.F Altschul	Basic local alignment search tool	Journal of Molecular Biology, 1990	Biochemistry & Molecular Biology
9	36,988 (56,581)	K.J. Livak	Analysis of relative gene expression data…	Methods, 2001	Biochemistry & Molecular Biology
10	32,657 (37,653)	N. Saitou	The neighbor-joining method—A new …	Molecular Biology and Evolution, 1987	Biochemistry & Molecular Biology; Evolutionary Biology; Genetics & Heredity
11	30,032 (33,046)	Z. Otwinowski,	Processing of X-ray diffraction data collected…	Macromolecular Crystallography, 1997	Biochemistry & Molecular Biology
12	29,615 (34,235)	A.D. Beckead	Density-functional exchange-energy …	Physical Review A, 1988	Physics
13	25,987 (29,094)	J.D. Thompson,	The CLUSTAL_X windows interface: flexible…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
14	25,880 (33,287)	R.M. Baron	The moderator mediator variable distinction…	Journal of Personality and Social Psychology, 1986	Psychology
15	25,696 (29,809)	J.M. Bland	Statistical methods for assessing agreement…	Lancet, 1986	General & Internal Medicine
16	25,340 (30,673)	T. Mosmann	Rapid colorimetric assay for cellular …	Journal of Immunological Methods, 1983	Biochemistry & Molecular Biology; Immunology
17	24,308 (28,923)	S. Iijima	Helical microtubules of graphitic carbon	Nature, 1991	Science & Technology - Other Topics
18	23,894 (34,400)	G. Kresse	Efficient iterative schemes for ab initio total-energy calculations using …	Physical Review B, 1996	Physics
19	23,294 (27,062)	J. Felsenstein	Confidence-limits on phylogenies – an approach using the bootstrap	Evolution, 1985	Environmental Sciences & Ecology; Evolutionary Biology; Genetics & Heredity
20	21,456 (21,529)	A.P. Feinberg	A technique for radiolabeling DNA restriction endonuclease fragments …	Analytical Biochemistry, 1983	Biochemistry & Molecular Biology; Chemistry

Table 1. Top 20 papers of citation.

3.2. Stochastic models

Simon [15] proposed the stochastic model, the so-called Simon's model, to elucidate the empirical distributions: distribution of words in prose samples by their frequency of occurrence, distributions of scientists by number of papers published, distributions of cities by population, distributions of income by size, and distributions of biological genera by number of species. Although assumptions of Simon's model are written in terms of word frequencies, we can express them in terms of network science as follows: assumption I—The probability that a node gets new link is proportional to the number of its degrees, that is, rich get richer or Matthew effect (e.g., see [16]), and assumption II—We add a new node with a constant probability γ. Simon's model elucidates the fact that the right-tail part of the distribution follows the power law distribution with $\mu = 1/(1-\gamma)$.

Price [17] generalized Simon's model, the so-called Price's model, to explain the growth of the citation networks. Barabáshi and Albert [18] introduced the stochastic model, the so-called BA model, based on two concepts: preferential attachment and growth, which corresponds to assumptions I and II of Simon's model, respectively. BA model is the case of $\gamma = 1/2$ of Simon's model and derives the power law distribution with $\mu = 2$. Jeong et al. [19] extended BA model to include an aging effect and a class of homogeneous connection kernels. Golosovsky and Solomon [20, 21] further extended to include an effect of initial attractivity.

Here, we use the model proposed by Jeong et al. [19] and check the aging effect and homogeneity of the growth of citation network. If we denote the number of degree of node i as k_i, the time evolution of k_i is obtained by

$$\frac{dk_i}{dt} = A_i(t)\, k_i^{\alpha}. \tag{2}$$

Here, $A_i(t)$ is an aging factor and $\alpha > 0$ is an unknown scaling exponent. Krapivsky et al. [22] have shown, for the case without the aging factor, for $\alpha = 1$ (linear preferential attachment) the model is just same as BA model and derives the power law distribution with $\mu = 2$. For $\alpha < 1$, the model derives the stretched exponential distribution, and for $\alpha > 1$ (super preferential attachment) a single node connects to nearly all other nodes, akin to gelation.

If we discretize the model and consider $\Delta t = 1$ year, Eq. (2) is written by

$$\Delta k_i = A_i k_i^{\alpha}, \tag{3}$$

We investigate the dynamics of growth for 44,932 papers published in 1985. The left panel of **Figure 4** depicts the double-logarithmic scale scatter plot of the number of citations, k_i ($i = 1, 2, \ldots, 44932$), as of 1988 and the change of the number of citations, Δk_i, from 1988 to 1999. If we divide k_i into bins with logarithmically equal separation, $\bar{k}$ and calculate the average value of Δk_i for each bin, $\bar{k}$, we obtain the red dots which are depicted in the right pane of **Figure 4**. By these manipulations, Eq. (3) is written by

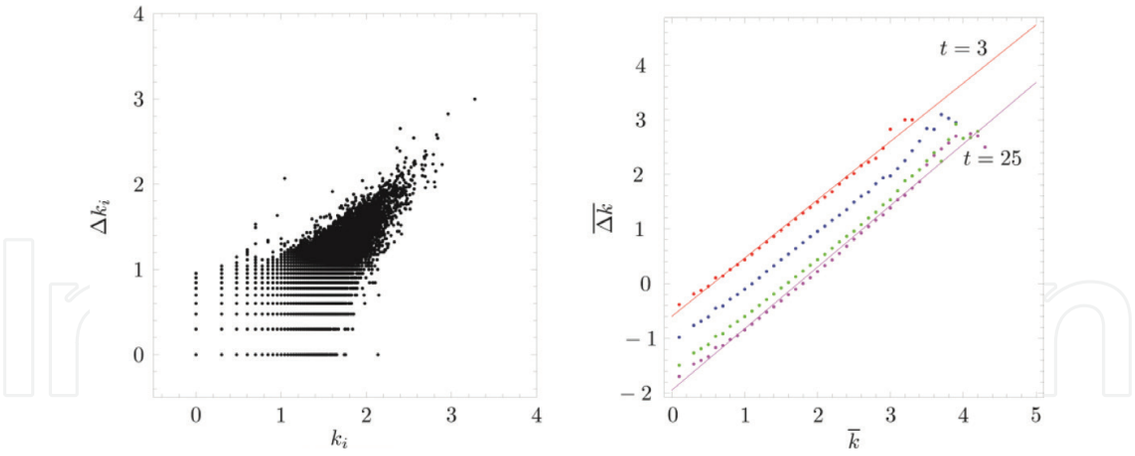

Figure 4. Left: Correlation between the number of citations and increase of the number of citations. Right: Change of the relation between mean citation and mean difference of citation.

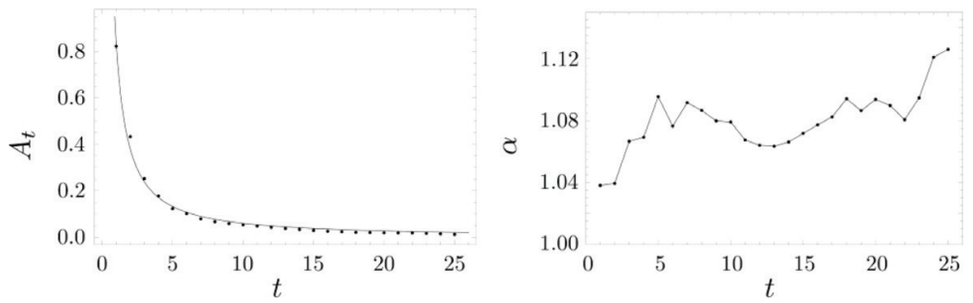

Figure 5. Left: Change of the aging effect. Right: Change of homogeneous factor.

$$\overline{\Delta k} = A_t \overline{k}^{\alpha}. \tag{4}$$

The red and solid line in the right panel of **Figure 4** corresponds to the linear regression of red dots by Eq. (4). The slope of this line corresponds to α and the intercept of it corresponds to A_t. In **Figure 4**, blue, green, and magenta dots are analysis for the year 1993, 2003, and 2010, respectively.

The left panel of **Figure 5** depicts the change of A_t. The solid line in this figure corresponds to the regression by the power law function given by $A_t \propto t^{-1.15}$. The right panel of **Figure 5** depicts the change of α. This figure shows that $\alpha > 1$ for the entire period in which we investigated. From this analysis, we realize that the citation network has the characteristics of super preferential attachment; therefore, it is expected that a single node connects to nearly all other nodes. However, the aging effect prevents the citation network from an oligopolistic network.

4. Distribution of PageRank

Google's PageRank is proposed by Brin and Page [4]. The Google number, G_i, of paper i is defined by the recursion formula (from Chen et al. [23]):

http://dx.doi.org/10.5772/intechopen.77389

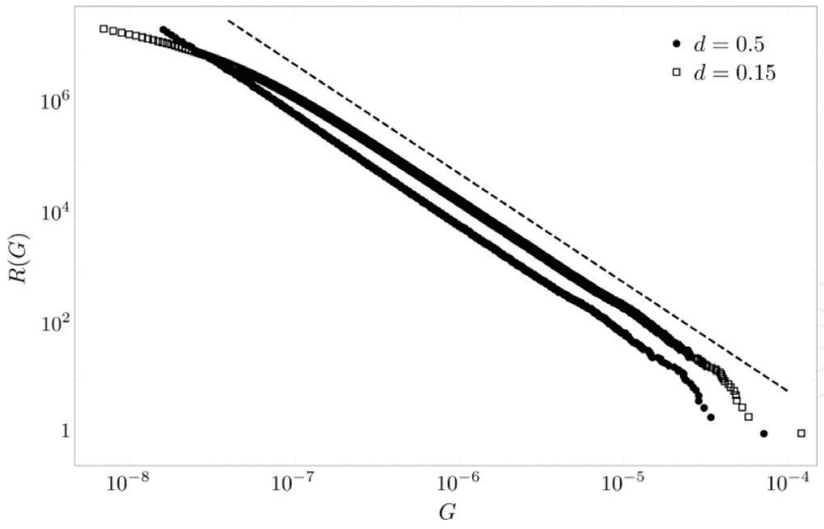

Figure 6. Rank size distribution, $R(G)$, of the Google number, G.

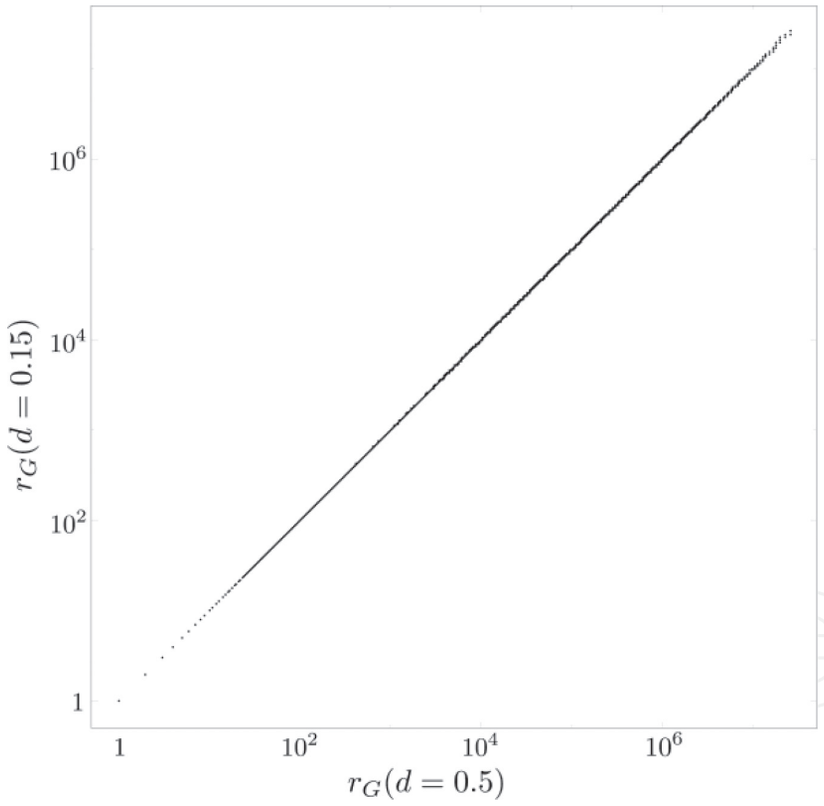

Figure 7. Correlation between the PageRank, r_G, in the case of $d = 0.5$ and $d = 0.15$.

$$G_i = (1-d)\sum_{i\,\mathrm{nn}\,j}\frac{G_j}{k_j}+\frac{d}{N}. \tag{5}$$

Here, $N = 34428322$ is the total number of articles contained in the largest connected component of the citation network. The sum is over the neighboring nodes j in which a link points to node i. In Eq. (5), d is a free parameter that controls the convergence and effectiveness

of the recursion calculation. In the original Google's PageRank [4], $d = 0.15$ is adopted and appropriate for the case of world wide web. On the other hand, $d = 0.5$ is adopted in [23] and appropriate for the case of citation network.

Figure 6 depicts the double-logarithmic scale plot of the rank size distribution of Google number, $R(G)$. In this figure, filled circles correspond to the case of $d = 0.5$ and open squares correspond to that of $d = 0.15$. The dashed line in this figure is the reference line and represents the power law distribution with $\mu = 2$. This value of exponent is same as the case of distribution of citation as depicted in **Figure 3**. Although the rank size distribution of Google number depends on d, the Google's PageRank, r_G, is almost the same as depicted in **Figure 7**. This figure is the double-logarithmic scale plot of r_G, and the abscissa is r_G in the case of $d = 0.5$, and the ordinate is r_G in the case of $d = 0.15$.

Table 2 depicts the top 20 lists of the Google's PageRank. The characteristics of this list are that papers belong to many subjects and that the publication years of papers are relatively old.

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
1	7.1314	4	45,349 (64,127)	4	G.M. Sheldrick	A short history of SHELX	Acta Crystallographica Section A, 2008	Chemistry; Crystallography
2	3.4074	1	60,967 (62,404)	0.5	P. Chomczynski	Single-step method of RNA isolation by acid…	Analytical Biochemistry, 1987	Biochemistry & Molecular Biology; Chemistry
3	3.1210	26	18,109 (18,789)	8.67	G.M. Sheldrick	Phase annealing in SHELX-90 – direct methods for…	Acta Crystallographica Section A, 1990	Chemistry; Crystallography
4	2.8852	2	55,143 (65,452)	0.5	A.D. Becke	Density-functional thermochemistry. 3…	Journal of Chemical Physics, 1993	Chemistry; Physics
5	2.8578	64	12,824 (14,640)	12.8	J. Kennedy	Particle swarm optimization	IEEE International Conference, 1995	Computer Science
6	2.7879	15	25,696 (29,809)	2.5	J.M. Bland	Statistical methods for assessing agreement…	Lancet, 1986	General & Internal Medicine
7	2.6547	3	52,035 (61,637)	0.43	C.T. Lee	Development of the Colle-Salvetti correlation…	Physical Review B, 1988	Physics
8	2.5745	76	11,685 (18,640)	9.5	D.G. Lowe	Distinctive image features from scale-invariant…	International Journal of computer Vision, 2004	Computer Science
9	2.4425	5	44,915 (64,682)	0.56	J.P Perdew	Generalized gradient approximation made…	Physical Review Letters, 1996	Physics

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
10	2.3890	46	14,128 (17,990)	4.6	S. Kirkpatrick	Optimization by simulated annealing	Science, 1983	Science & Technology - Other Topics
11	2.3430	11	30,032 (33,046)	1	Z. Otwinowski	Processing of X-ray diffraction data…	Macromolecular Crystallography, 1997	Biochemistry & Molecular Biology
12	2.3236	97	10,368 (11,590)	8.08	F.H. Allen	Table of bond lengths determined by X-RAY…	Journal of the Chemical Society-Perkin Transactions 2, 1987	Chemistry
13	2.2868	6	42,407 (46,286)	0.56	J.D. Thompson	Clustal-W - improving the sensitivity of…	Nucleic Acids Research, 1994	Biochemistry & Molecular Biology
14	2.1787	8	37,133 (48,832)	0.57	S.F. Altschul	Basic local alignment search tool	Journal of Molecular Biology, 1990	Biochemistry & Molecular Biology
15	2.1481	7	39,281 (44,765)	0.47	S.F. Altschul	Gapped BLAST and PSI-BLAST: a new…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
16	2.0319	10	32,657 (37,653)	0.63	N. Saitou	The neighbor-joining method – a new method…	Molecular Biology and Evolution, 1987	Biochemistry & Molecular Biology; Evolutionary Biology; Genetics & Heredity
17	1.9081	17	24,308 (28,923)	1	S. Iijima	Helical microtubules of graphitic carbon	Nature, 1991	Science & Technology - Other Topics
18	1.8685	107	9775 (10,827)	5.94	H.D. Flack	On enantiomorph-polarity estimation	Acta Crystallographica Section A, 1983	Chemistry; Crystallography
19	1.8001	82	11,242 (12,850)	4.32	A.L. Spek	Single-crystal structure validation with the…	Journal of Applied Crystallography, 2003	Chemistry; Crystallography
20	1.7796	129	8818 (8849)	6 s.45	N. Walker	An empirical-method for correcting…	Acta Crystallographica Section A, 1983	Chemistry; Crystallography

Table 2. Top 20 papers of Google's PageRank.

5. Correlation between citation and PageRank

Bollen and Rodriquez [24] described that the Institute for Scientific Information (ISI) Impact factor (IF) which is defined as the mean number of citations a journal receives over a two-year

period is a metric of popularity and that the Google's PageRank is a metric of prestige. This concept is also proposed by Chen et al. [23] and Maslov and Redner [25] which investigated all publications in the Physical Review family of journals from 1893 to 2003 and found the linear relation between the Google number and the number of citations. Furthermore, [23, 25] found that some outliers from this linear relation, especially the papers of which the ranking of PageRank is remarkably high and that of citation is slightly high, are universally familiar to physicists [23, 25] called such papers scientific "gems." Ma et al. [26] applied the concept of [23–25] to the field of biochemistry and molecular biology from 2000 to 2005. Though these studies investigated the citation network of some selected scientific field, this chapter investigates the citation network consisting of all scientific fields.

Figure 8 depicts the double-logarithmic scale plot of the correlation between the number of citations, k, and the Google number, G. In this figure, the solid gray line represents the mean value $\langle G \rangle$ calculated for bins of k with logarithmically equal width. This figure shows that $\langle G \rangle$ versus k is smooth and increases linearly with k for $k \geq 500$. Thus, the Google number and citations are almost similar measures characterizing the importance of papers. This result means that prestige (Google number) is proportional to popularity (citations) in many cases.

However, there are outliers which have high prestige comparing to popularity. These papers are located above the solid gray line in **Figure 8** and are regarded as extremely prestigious papers. If we denote the citation rank as r_k and the Google's PageRank as r_G, these extremely prestigious papers are extracted by the order of Google's PageRank with the constraint given by the ratio r_k/r_G. **Table 3** depicts the top 20 extremely prestigious papers selected by using the constraint $r_k/r_G > 10$. The characteristic of this list is that the subjects of papers are almost information science.

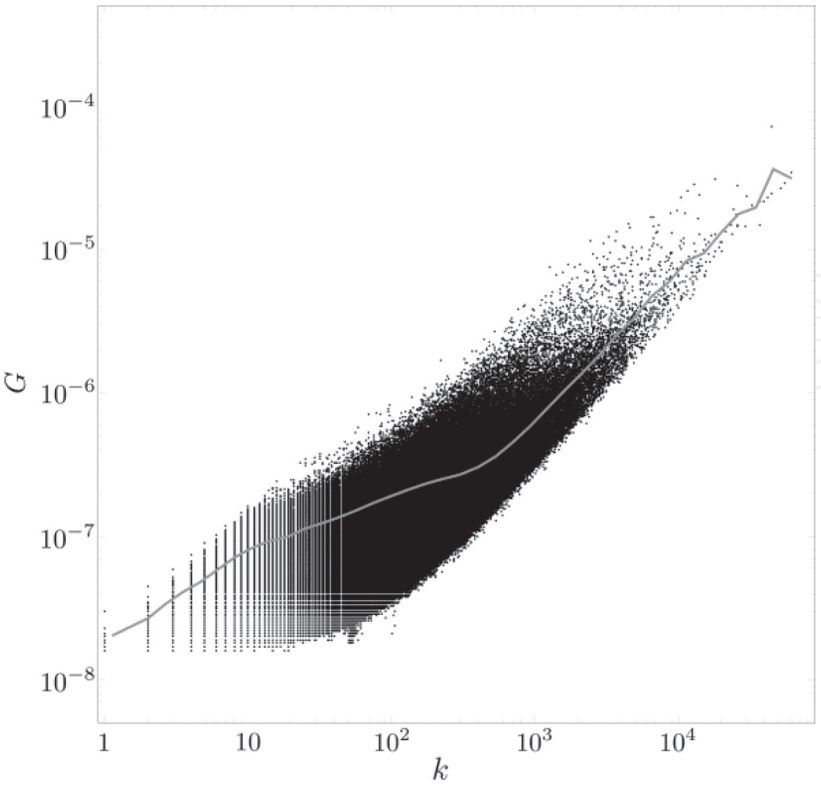

Figure 8. Correlation between the number of citations, k, and the Google number, G.

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
5	2.8578	64	12,824 (14,640)	12.8	J. Kennedy	Particle swarm optimization	Proceedings of IEEE International Conference, 1995	Computer Science
22	1.6861	240	6500 (7458)	10.91	S.M. Alamouti	A simple transmit diversity technique for wireless…	IEEE Journal on Selected Areas in Communications, 1998	Engineering; Telecommunications
25	1.5103	516	4465 (6605)	20.64	I.F. Akyildiz	Wireless sensor networks: a survey	Computer Networks, 2002	Computer Science; Engineering; Telecommunications
33	1.4160	481	4611 (6276)	14.58	Z. Pawlak	Rough sets	International Journal of Computer & Information Sciences, 1982	Information Science & Library Science
36	1.3169	784	3740 (5402)	21.78	I.F. Akyildiz	A survey on sensor networks	IEEE Communications Magazine, 2002	Engineering; Telecommunications
43	1.2155	998	3309 (4707)	23.21	T.R. Gruber	A translation approach to portable ontology…	Knowledge Acquisition, 1993	Computer Science; Information Science & Library Science
48	1.1432	828	3656 (4463)	17.25	P. Gupta	The capacity of wireless networks	IEEE Transactions on Information Theory, 2000	Computer Science; Engineering
49	1.1387	1916	2441 (2839)	39.10	S. Floyd	Random early detection gateways for congestion…	IEEE-ACM Transactions on Networking, 1993	Computer Science; Engineering; Telecommunications
53	1.1102	1247	2991 (3879)	23.53	G. Bianchi	Performance analysis of the IEEE 802.11 distributed…	IEEE Journal on Selected Areas in Communications, 2000	Engineering; Telecommunications
60	1.0626	608	4149 (5968)	10.13	S. Haykin	Cognitive radio: Brain-empowered wireless…	IEEE Journal on Selected Areas in Communications, 2005	Engineering; Telecommunications
76	0.9431	967	3360 (3961)	12.72	T. Murata	Petri nets - properties, analysis and applications	Proceedings of the IEEE, 1989	Engineering
79	0.9388	1758	2535 (3702)	22.25	W.B. Heinzel-man	An application-specific protocol architecture for…	IEEE Transactions on Wireless Communications, 2002	Engineering; Telecommunications
90	0.8884	1190	3048 (4075)	13.22	R. Ahlswede	Network information flow	IEEE Transactions on Information Theory, 2000	Computer Science; Engineering

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
93	0.8767	1565	2691 (3401)	16.83	T. Wiegand	Overview of the H.264/AVC video coding standard	IEEE Transactions on Circuits and Systems for Video Technology, 2003	Engineering
97	0.8598	1045	3245 (4674)	10.77	M. Dorigo	Ant system: Optimization by a colony of…	IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 1996	Automation & Control Systems; Computer Science
116	0.7923	2736	2052 (2426)	23.59	D. HAREL	Statecharts - a visual formalism for…	Science of Computer Programming, 1987	Computer Science
120	0.7838	4059	1705 (2982)	33.83	M. WEISER	The Computer for the 21st-century	Scientific American, 1991	Science & Technology - Other Topics
121	0.7796	1406	2840 (4011)	11.62	S. Deerwester;	Indexing by latent semantic analysis	Journal of the American Society for Information Science, 1990	Computer Science; Information Science & Library Science
128	0.7584	3165	1914 (1948)	24.73	A.E. Leviton	Standards in herpetology and ichthyology…	Copeia, 1985	Zoology
129	0.7582	7409	1274 (1478)	57.43	X.Y. Wang	Room-temperature all-semiconduc-ting…	Physical Review Letters, 2008	Physics

Table 3. Top 20 extremely prestigious papers.

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
125	8890 (17,192)	627	0.3250	5.02	D. Hanahan	Hallmarks of Cancer: The Next Generation	Cell, 2011	Biochemistry & Molecular Biology; Cell Biology
297	5817 (10,877)	1580	0.2042	5.32	D.W. Huang	Systematic and integrative analysis of large gene list…	Nature Protocols, 2008	Biochemistry & Molecular Biology
304	5747 (9681)	1608	0.2023	5.29	Y. Zhao	The M06 suite of density functionals for main…	Theoretical Chemistry Accounts, 2008	Chemistry
327	5533 (8874)	1810	0.1897	5.54	D.P. Bartel	MicroRNAs: Target Recognition and…	Cell, 2009	Biochemistry & Molecular Biology; Cell Biology

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
375	5147 (6894)	2128	0.1757	5.67	B.P. Lewis	Conserved seed pairing, often flanked by…	Cell, 2005	Biochemistry & Molecular Biology; Cell Biology
414	4912 (5825)	2506	0.1619	6.05	T. Jenuwein	Translating the histone code	Science 2001	Science & Technology - Other Topics
419	4895 (5350)	2123	0.1759	5.07	P. Li	Cytochrome c and dATP-dependent formation…	Cell, 1997	Biochemistry & Molecular Biology; Cell Biology
535	4382 (4604)	2802	0.1534	5.24	Z.G. XIA	Opposing effects of ERK and JNK-P38 map…	Science, 1995	Science & Technology - Other Topics
543	4343 (5864)	3120	0.1447	5.75	R.C. LEE	The C. elegans heterochronic geneG…	Cell, 1993	Biochemistry & Molecular Biology; Cell Biology
547	4327 (4633)	2865	0.1517	5.24	A. Hall	Rho GTPases and the actin cytoskeleton	Science, 1998	Science & Technology - Other Topics
600	4164 (5479)	3269	0.1411	5.45	S. Akira	Pathogen recognition and innate immunity	Cell, 2006	Biochemistry & Molecular Biology; Cell Biology
611	4144 (4888)	3585	0.1348	5.87	B.D. Strahl	The language of covalent histone modifications	Nature, 2000	Science & Technology - Other Topics
640	4063 (5604)	3359	0.1390	5.25	M.E. Raichle	A default mode of brain function	PNAS, 2001	Science & Technology - Other Topics
645	4054 (5303)	3326	0.1398	5.16	E.K. Miller	An integrative theory of prefrontal cortex function	Annual Review of Neuroscience, 2001	Neurosciences & Neurology
657	4026 (4967)	3572	0.1351	5.44	R.O. Hynes	Integrins: Bidirectional, allosteric signaling…	Cell, 2002	Biochemistry & Molecular Biology; Cell Biology
661	4005 (4335)	4096	0.1262	6.20	S.R. Datta	Akt phosphorylation of BAD couples survival…	Cell, 1997	Biochemistry & Molecular Biology; Cell Biology
706	3912 (5288)	4825	0.1166	6.83	T. Kouzarides	Chromatin modifications and their function	Cell, 2007	Biochemistry & Molecular Biology; Cell Biology

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
751	3806 (5109)	4096	0.1262	5.45	M. Corbetta	Control of goal-directed and stimulus-driven…	Nature Reviews Neuroscience, 2002	Neurosciences & Neurology
752	3805 (4313)	4446	0.1213	5.91	A. Brunet	Akt promotes cell survival by phosphorylating and…	Cell, 1999	Biochemistry & Molecular Biology; Cell Biology
785	3739 (4363)	3972	0.1280	5.06	J.D. Fontenot	Foxp3 programs the development and…	Nature Immunology, 2003	Immunology

Table 4. Top 20 extremely popular papers.

On the other hand, there are also outliers which have low prestige comparing to popularity. These articles are located below the solid gray line in **Figure 8** and are regarded as extremely popular papers. These articles are extracted by the order of citation rank with the constraint given by the ratio r_G/r_k. **Table 4** depicts the top 20 extremely popular papers selected by using the constraint $r_G/r_k > 5$. These articles are divided into two groups. One group contains papers which are published in Nature, Science, and the Proceedings of the National Academy of Science of the United State of America (PNAS). Besides, publication year of these papers are approximately over 10 years ago. Furthermore, the growth rate of citations, k'/k, of those papers are low. The other group includes papers which are mainly published in Cell and are published relatively recently. What is more, the growth rate of citations, k'/k, of those papers are extremely high. Thus, we can regard these papers as rising papers.

6. Conclusions

We investigated papers published from 1981 to 2015 and contained in SCI-E. The total number of papers is 34,666,719 and that of citations is 591,321,826. We extracted the largest connected component from this dataset. The obtained citation network consists of 34,428,322 nodes (articles) and 591,177,607 links (citations).

The right-tail part of the rank size distribution of citations follows the power law distribution with exponent $\mu = 2$, that is, $R(k) \propto k^{-2}$. Furthermore, we introduced the generalized beta distribution of the second kind (GB2) as the best-fit function to the whole range of citation distribution. We introduced the stochastic model with growth, preferential attachment, and aging effect. Through the numerical analysis, we obtained the value of the parameter set.

Although the number of citations represent the popularity of papers, Google's PageRank reflects the prestige of papers. We evaluated Google's PageRank for the largest connected component which consists of 34,428,322 articles and 591,177,607 link citations. We found that the citations and Google numbers have a positive linear relation. We consider this positive

linear relation as a benchmark and selected extremely prestigious and extremely popular papers. We found that the subject of extremely prestigious papers is almost information science. Furthermore, we found that extremely popular papers are divided into popular papers and rising papers.

We conclude this chapter by describing two remaining issues. One concerns the stochastic model. Though we introduce GB2 as the best-fit function to the whole range of citation distribution, there is no stochastic model that explains GB2. The other concerns the weight of links in the citation network. Almost all studies have investigated citation networks as unweighted networks. However, it is possible to define weight of links, for example, similarity between papers.

Acknowledgements

This work is supported by Nihon University College of Science and Technology Grants-in Aid 2012 and 2016. The authors thank the Yukawa Institute of Theoretical Physics at Kyoto University. Discussions during the YITP workshop YITP-W-17-14 on ”Econophysics 2017“ were useful to complete this work.

Author details

Wataru Souma[1]* and Mari Jibu[2]

*Address all correspondence to: souma.wataru@nihon-u.ac.jp

1 College of Science and Technology, Nihon University, Japan

2 Japan Science and Technology Agency, Japan

References

[1] Hou J. Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy. Scientometrics. 2017;**110**:1437-1452. DOI: 10.1007/s11192-016-2206-9

[2] de Solla Price DJ. Networks of Scientific Papers. Science. 1965;**149**:510-515. DOI: 10.2307/1716232

[3] Bollen J, Van de Sompel H, Hagberg A, Chute R. A principal component analysis of 39 scientific impact measures. PLoS One. 2009;**4**:e6022. DOI: 10.1371/journal.pone.0006022

[4] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998;**30**:107-117. DOI: 10.1016/S0169-7552(98)00110-X

[5] Pareto V. Cours d'économie politique: professé a l'université de lausanne - tone second. Rouge: Lausanne F; 1897

[6] Arnold BC. Pareto Distributions. 2nd ed. US: CRC Press; 2015. p. 456. ISBN: 9781466584846

[7] Aoyama H, Fujiwara Y, Ikeda Y, Iyetomi H, Souma W, Yoshikawa H. Macro-Econophysics: New Studies on Economics Networks and Synchronization. UK: Cambridge University Press; 2017. pp. 53-96. ISBN: 9781107198951

[8] Rednar S. How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B. 1998;**4**:131-134. DOI: 10.1007/s100510050359

[9] Redner S. Citation statistics from 110 years of physical review. Physics Today. 2005;**58**: 49-54. DOI: 10.1063/1.1996475

[10] Albarrán P, Ruiz-Castillo J. References made and citations received by scientific articles. Journal of the American Society for Information Science and Technology. 2011;**62**:40-49. DOI: 10.1002/asi.21448

[11] Albarrán P, Crespo JA, Ortuño I, Ruiz-Castillo J. The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics. 2011;**88**:385-397. DOI: 10.1007/s11192-011-0407-9

[12] Brzezinski M. Power laws in citation distributions: Evidence from Scopus. Scientometrics. 2015;**103**:213-228. DOI: 10.1007/s11192-014-1524-z

[13] McDonald JB. Some generalized functions for the size distribution of income. Econometrica. 1984;**52**:647-663. DOI: 10.2307/1913469

[14] Kleiber C, Kotz S. Macro-Econophysics: Statistical Size Distributions in Economics and Actuarial Sciences. John Wiley and Sons; 2003. DOI: 10.1002/0471457175.ch2

[15] Simon HA. On a class of skew distribution functions. Biometrika. 1955;**42**:425-440. DOI: 10.2307/2333389

[16] Merton RK. The Matthew effect in science. Science. 1968;**159**:56-63. DOI: 10.1126/science.159.3810.56

[17] de Solla Price DJ. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science. 1976;**27**:292-306. DOI: 10.1002/asi.4630270505

[18] Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;**286**: 509-512. DOI: 10.1126/science.286.5439.509

[19] Jeong H, Néda Z, Barabási AL. Measuring preferential attachment in evolving networks. Europhysics Letters. 2003;**61**:567-572. DOI: 10.1209/epl/i2003-00166-9

[20] Golosovsky M, Solomon S. Stochastic dynamical model of a growing citation network Based on a Self-Exciting Point Process. Physical Review Letters. 2012;**109**:098701. DOI: 10.1103/PhysRevLett.109.098701

[21] Golosovsky M, Solomon S. Growing complex network of citations of scientific papers: Modeling and measurements. Physical Review E. 2017;**95**:012324. DOI: 10.1103/PhysRevE.95.012324

[22] Krapivsky P, Redner S, Leyvraz F. Connectivity of Growing Random Networks. Physical Review Letters. 2000;**85**:4629-4632. DOI: 10.1103/PhysRevLett.85.4629

[23] Chen P, Xie H, Maslov S, Redner S. Google PageRank algorithm, scientific gems, physical review Citations. Journal of Informetrics. 2007;**1**:8-15. DOI: 10.1016/j.joi.2006.06.001

[24] Bollen J, Rodriquez MA, Van de Sompel H. Journal status. Scientometrics. 2006;**69**: 669-687. DOI: 10.1007/s11192-006-0176-z

[25] Maslov S, Redner S. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. Society for Neuroscience. 2008;**28**:11103-11105. DOI: 10.1523/JNEUROSCI.0002-08.2008

[26] Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Information Processing & Management. 2008;**44**:800-810. DOI: 10.1016/j.ipm.2007.06.006

IntechOpen