Technological Forecasting & Social Change 81 (2014) 49–55

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Predicting highly cited papers: A Method for Early Detection of Candidate Breakthroughs

Ilya V. Ponomarev [a,*], Duane E. Williams [a], Charles J. Hackett [b], Joshua D. Schnell [a], Laurel L. Haak [a]

[a] Thomson Reuters, Rockville, MD, USA
[b] Division of Allergy, Immunology, and Transplantation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 16 March 2012
Received in revised form 30 July 2012
Accepted 26 September 2012
Available online 21 November 2012

Keywords:
Bibliometrics
Scientometrics
Highly cited papers
Breakthrough paper indicator
Research management
Technological forecasting
Science policy

ABSTRACT

Scientific breakthroughs are rare events, and usually recognized retrospectively. We developed methods for early detection of candidate breakthroughs, based on dynamics of publication citations and used a quantitative approach to identify typical citation patterns of known breakthrough papers and a larger group of highly cited papers. Based on these analyses, we proposed two forecasting models that were validated using statistical methods to derive confidence levels. These findings can be used to inform research portfolio management practices.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Understanding the process of scientific knowledge creation and measuring impact is vital from both a management and policy standpoint. Scientific discoveries may be gradual and incremental, or transformative and abrupt. While both paths to discovery are important, major advances in science strongly depend upon explosive breakthrough discoveries. By detecting these transformative discoveries as early as possible, more time is gained to support these emerging research areas through workshops, new funding, or collaborative research efforts.

The quantitative study of ground breaking research publications has a long history, and has been based typically on publication citation statistics [1–3]. In particular, de Solla Price [4] showed that the citation count distribution for individual publications had a power-law form. He noted that well-cited papers continued to be referenced more frequently than less-cited papers, and coined [5] the term “cumulative advantage” [6] to describe the mechanism that causes a persistently higher rate. In the framework of network models, this mechanism is now known as preferential attachment [7]. In the 1980s, Pendlebury [8] and Garfield [9] performed analysis of total citations count of eminent researchers and showed [10] that most-cited author ranking effectively identified Nobel Prize winners. This approach is used currently in the *Essential Science Indicators*^{SM} (ESI) product from Thomson Reuters [11]. Aversa [12] identified 2 different citation patterns for 400 highly cited papers published in the 1970s and analyzed aging rates for those papers. Growth and aging effects were also analyzed and modeled by van Raan [13]. More recently, Redner [14] analyzed citation statistics using a corpus of all papers published in Physics Review Journals during its 110 year history. He demonstrated temporal features associated with citations, such as citation patterns, highly correlated burst of citations, and downturns in research activity. Chen et al. [15] introduced the explanatory and computational theory of transformative discovery based on a network approach to scientific knowledge diffusion.

[*] Corresponding author at: 1455 Research Blvd., 2nd Floor, Rockville MD 20850, USA. Tel.: +1 301 545 4259.
E-mail address: ilya.ponomarev@thomsonreuters.com (I.V. Ponomarev).

0040-1625/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.techfore.2012.09.017

The present work pursues a different goal: to describe and to validate a scalable method for early identification of breakthrough candidate publications (BPs) by predicting future citation patterns of individual papers in a collection using time dependent analysis of citation rates. It is worth noting that, without a doubt, identifying influential discoveries is a multidimensional process like that of research itself and should involve metrics beyond simple cumulative citation counts. Examples include ranking citations by geographic region, by interdisciplinary features [16–18], by prestige diversity, recognition by leading experts, by count and classifications of awards received, media coverage, and by informal citations (names in titles, acknowledged methods abbreviations etc.). This manuscript is a first step in developing such a multidimensional breakthrough paper indicator.

2. Method and analysis

In this paper we focus on ranked citation counts and monthly citation rates as proxies for scientific impact. Our approach can be described in the following steps:

1. Consider a small set of known BPs in a particular research field.
2. For each BP identify a statistically large set of similar publications.
3. Rank by cumulative citations, establish and justify citation breakthrough thresholds.
4. Identify typical pattern of time-dependent citation behavior of highly cited papers.
5. Choose theoretical model which describes this behavior with high level of statistical confidence.
6. Using only knowledge about earlier times of citation curves (6–24 months) fit and interpolate results on later times.
7. Compare and validate predicted results empirically with actual citation behavior.

2.1. Citation data

For our studies we used citation data sets derived from Thomson Reuters Web of Science (WOS). Citation data were extracted in January, 2012. Firstly, as a reference, an initial set of 11 known breakthrough publications in molecular biology and genetics was compiled by subject matter experts in the Division of Allergy, Immunology, and Transplantation (DAIT) of the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health [19–29]. In our analysis we abbreviate this set as "DAIT BPs" (see Table 1).

For each DAIT paper we identified a set of similar papers. We considered a paper similar if it was published in the same research field and during the same calendar year. Research field selection was based on ESI. Each journal was assigned to one of the 22 major fields of sciences. Each paper was then assigned to a discipline—and only one discipline—based on the journal in which it appeared. Such selection allows us to generate a data corpus that supported statistical significance testing. General information about DAIT BPs and corresponding similar paper sets is given in Table 1.

Furthermore, to estimate breakthrough citation thresholds in all 22 ESI subject categories we evaluated annual data sets of research articles, published from 1995 to 2005, that acknowledged NIH funding support. We call these data sets "MEDLINE Sets". We validated our findings using the 2005 data set (total 375,372 items). Citation data were analyzed in one-month intervals. A time stamp of citation was determined by journal issue date of citing publication. In the rare events when journal issue date was not available it was calculated by using the known journal issue frequency.

Table 1
DAIT BPs citation statistics. '1st Author Name' — 1st author names in referenced set [15–25]; 'Subject Category' — WOS subject category (MD is for multidisciplinary, and MBG is for Molecular Biology and Genetics); 'Cites 6M/24/60' — number of citations at 6/24/60 months since publication; 'JCBI 24' - Journal Citation Benchmark Indicator (JCBI) — median value of citations after 24 months for papers published in the journal; 'Cites Ratio' — ratio between 'JCBI 24' and 'Cites 24'; '# Sim Pubs' — number of similar publications in the comparison set; 'Paper Rank' — citation rank of the paper in the set (#1 corresponds to the highest cited paper); 'Top %' — percentile position of the article in the ranked list; 'In Top 1%' — whether or not (Y/N) paper made a top 1% cut; and 'Cites Now' — total number of citations received as of January 2012.

1st Author name	Journal	Subject category	Pub year	Cites 6 M	Cites 24 M	JCBI 24	Cites Ratio	# Sim pubs	Cites 60 M	Paper rank	Top %	In top 1%	Cites now
Hammond	Nature	MD	2000	15	146	20	7.3	5105	603	38	0.74%	Y	1317
Ketting	Nature	MD	2000	4	40	20	2	5105	87	1269	24.8%	N	116
Domeier	Science	MD	2000	10	47	25	1.9	5105	75	1493	29.3%	N	106
Caplen	Gene	MBG	2000	3	33	3	11	33,209	92	1885	5.68%	N	131
Lagos-Quintana	Science	MD	2001	26	160	27	5.9	5135	491	68	1.32%	N	1267
Fire	Nature	MD	1998	11	207	26	8	5299	995	13	0.25%	Y	4856
Distel	Cell	MBG	1987	3	129	31	4.2	18,566	325	62	0.33%	Y	498
McHeyzer-Williams	Science	MD	1995	0	59	22	2.7	4898	165	485	9.90%	N	351
Hicke	Cell	MBG	1996	7	107	43	2.5	29,663	250	238	0.80%	Y	513
Nussenzweig	Nature	MD	1996	6	67	28	2.4	5152	203	309	6.0%	N	424
Altman	Science	MD	1996	5	81	21	3.9	5152	699	27	0.52%	Y	2236

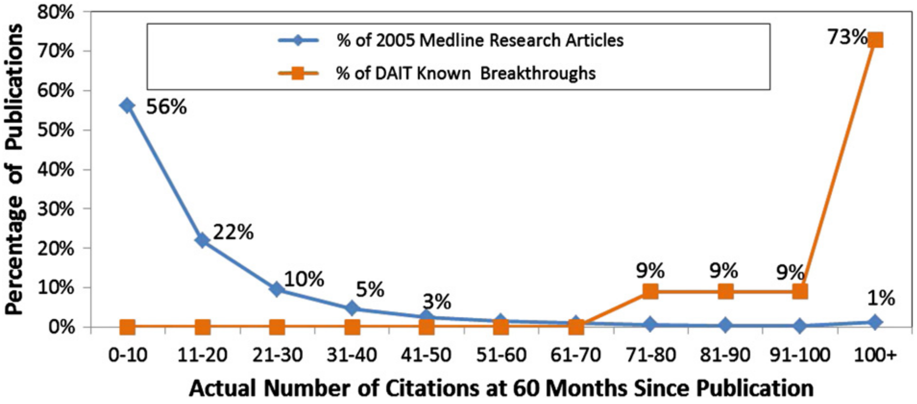

Fig. 1. Distribution of cumulative number of citations after 5 y for two data sets. MEDLINE set in the 2005 data set (blue, 375 K articles) show typical highly skewed power law behavior with only about 1% of the papers being cited more than 100 times. In comparison, all DAIT BPs (orange) have been cited more than 75 times and 8 papers acquired more than 100 citations.

2.2. Citation breakthrough threshold

Our initial retrospective analysis was performed using a 2005 MEDLINE data set. We established a 5 year time window after publication date for a calculation of cumulative citation rankings, assuming that this time interval is sufficient for recognition of seminal papers by the scientific community. Our criterion for a candidate breakthrough was a paper that exceeded a certain threshold of cumulative citations count 5 years after publication. A comparison of cumulative citation distributions between DAIT and 2005 MEDLINE set is shown in Fig. 1.

Thus an important question was the determination of the numeric value for the breakthrough citation threshold. A simple assignment of the same high value of citations (e.g., 100) is not feasible, as it does not account for variation in citation behavior between fields. As shown in Fig. 2, a threshold of 100 citations would lead to 1858 BP candidates in Clinical Medicine in 2005 alone, and more than 500 papers in four other subject categories. Such a high rate of BPs does not correspond with the intuitive expectation regarding rareness of breakthrough discoveries. On the other hand, using this threshold, some categories like Space Science or Economics would have no breakthrough candidates. The reason for such divergence is that the number of citations depends strongly on publication volumes and on the size of a particular scientific community, which vary substantially by *field of research* (see right column in Fig. 2).

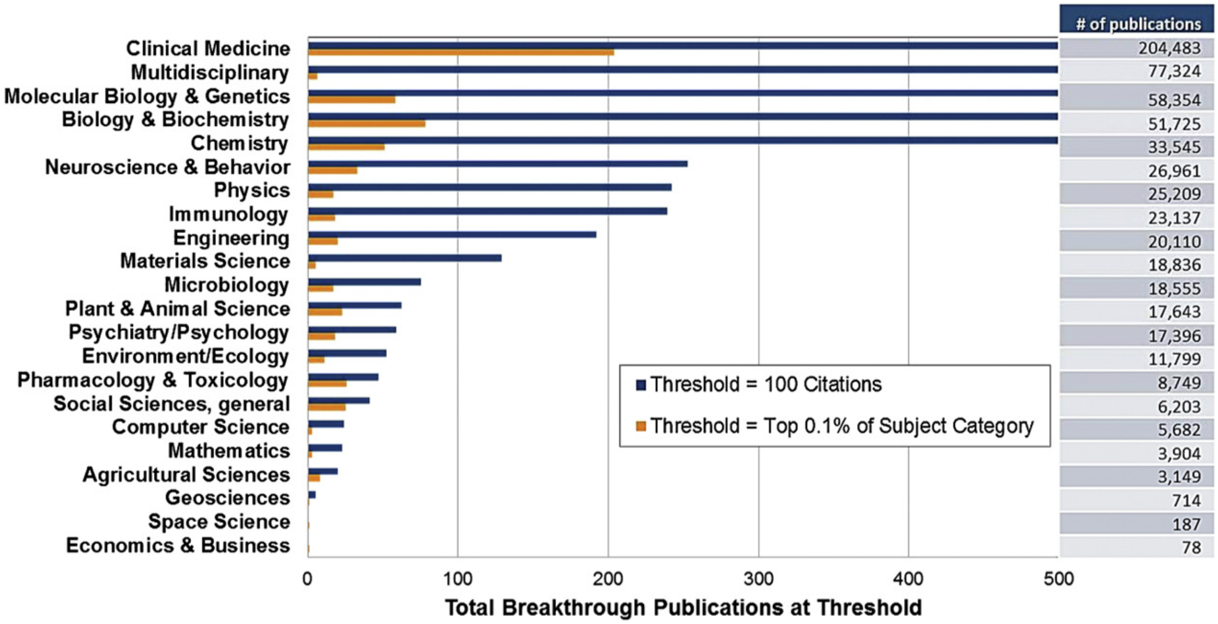

Fig. 2. BPs citation thresholds and corresponding number of papers, by subject category for numeric (blue) and percentile (orange) threshold selection.

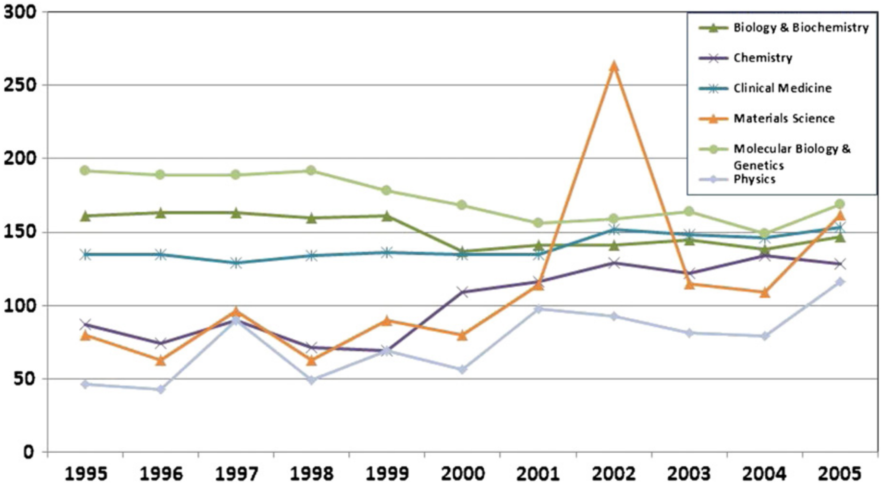

Fig. 3. Paper count at 0.1% citation threshold by year for several select fields of study.

We found that a *percentile approach* for establishing the threshold value is more effective than a strict numerical cutoff. In general, the *top percentage* of publications in a set should be selected as a cutoff point. Selectivity with inclusiveness needs to be balanced so as to choose a selective group for future monitoring, and then apply other metrics to further filter the results. We determined that for topical sets containing more than 20,000 publications, a 0.1% cutoff is optimal (Figs. 2 and 3). For smaller sets, it may be desirable to increase the cutoff to 1% or more.

We also studied MEDLINE annual data sets for each of the years 1995–2004 to test whether the citation threshold changes with time (see Fig. 3). While some of the subject categories have monotonic dependencies or strong fluctuations, we found that research fields related to medicine and biology have a stable citation threshold. This is a very important finding since it allows us to project current average value of citation threshold in future years.

2.3. Identifying typical citation patterns

The next step of our detection method was to identify citation patterns of top-cited publications using monthly citation counts and rates. In Fig. 4, the cumulative citation count (left panel) and its derivative – a monthly citation rate – (right panel) are shown. In general, a typical citation pattern has an initial period of slow citation growth that lasts from 5 to 20 months (monthly rate is proportional t^{a} with $1 \leq \alpha \leq 2$). Following this initial slow growth phase, the citation rates accelerate. Then citation rates reach saturation plateaus, after which they decrease (memory or aging effect). While for a majority of top ranked publication citations count will follow this scenario, the time transitions between these phases vary substantially from paper to paper. We analyzed the top cited papers in several datasets. We found that after 5 years, approximately 25% are still in the first stage of growth (Type A, right panel of Fig. 4), 50% are in saturation (Type B) and 25% have started aging (Type C).

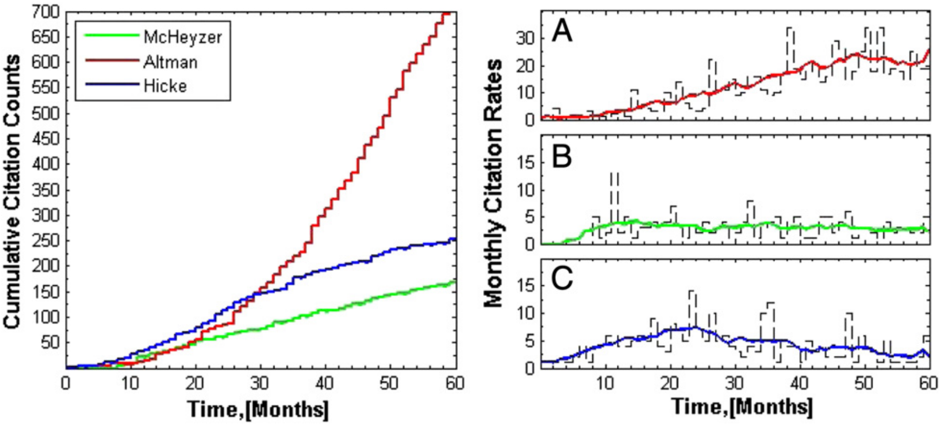

Fig. 4. Three typical citation count patterns. Due to strong fluctuations, monthly citation rates were smoothed by applying a moving average window with a 5 month span. Please note the different upper limits for y-axis for monthly citation rates.

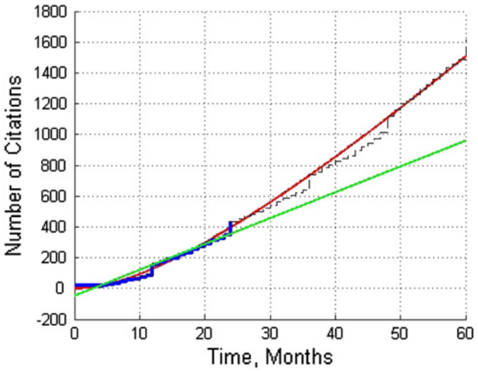

Fig. 5. Citation curves of the actual paper (dashed line), linear (green) and nonlinear (red) fitting curves. The fitting procedure used only citation data for the first several months (shown in blue). Data were extrapolated to predict future citations and compared with actual citation curve.

2.4. Forecasting models

Based on these observations, we developed a forecasting model that allowed us to predict most citation trajectories with a minimum number of fitting parameters. Our approach was the following: using data for the first n months (n = 6, 12, and 24), we curve-fit citation data and extrapolated results to time t = 5 years and predicted whether or not the paper had breakthrough potential (e.g., the citation threshold will be above what is expected for a given subject category).

In our method, we chose linear model (LM) f_1 and non-linear model (NLM) f_2 theoretical curves to fit the citation behavior (Fig. 5):

$$f_1(t) = a_1 t + b_1,$$

$$f_2(t) = a_2 b_2 t - a_2 b_2^2 \log\left(1 + \frac{t}{b_2}\right).$$

3. Model validation and results

We used the 2005 MEDLINE BPs for the Chemistry Subject Category as a test set. This set comprises 169 journals and 51,575 publications. There were 51 papers above the 0.1% citation breakthrough threshold (262 citations after 5 years). We experimented with optimization of the initial detection time window (i.e., the initial time span in which citation data necessary for forecasting model input are collected). Optimization of this window is driven by a tradeoff between three contradicting requirements. First, from the perspective of science policy makers, this window should be as short as possible to allow BPs to be detected as early as possible. Second, for a fitting process the more data (citations events) you have, the better the fit quality, and thus there needs to be a minimum number of data points to generate a reasonable data fit (e.g., if you have only three data points, varying different fitting parameters will result in an infinite number of solutions through them, which renders the results meaningless). Third, the initial time window should be long enough to account for the editorial cycles of different journals. To be cited, a paper must be incorporated into a new manuscript that must then be submitted to a journal, peer-reviewed, revised, and then finally published. Choosing too short a window eliminates paper journals with longer editorial cycles from consideration. Therefore, the initial time window should be adjusted to be relevant for a statistically large enough group of journals. It should be noted that application of a minimum citation count cutoff at the early months inevitably risks omission of a group of papers which became highly cited later (aka "sleeping beauties" or "late bloomers" [30]) as well as those from journals with longer editorial cycles.

Table 2
Precision and recall for linear model.

	Months of data	6	12	24	36	60
1	# Papers predicted	1	5	13	22	27
2	# Correctly predicted	1	4	12	20	26
3	# Incorrectly predicted	0	1	1	2	1
4	Precision (row#2/row#1, %)	100	80	92	91	96
5	Recall (row #2/30 actual, %)	3	13	40	67	87

Table 3
Precision and recall for non-linear model.

	Months of data	6	12	24	36	60
1	# Papers predicted	266	51	34	32	29
2	# Correctly predicted	12	19	23	25	28
3	# Incorrectly predicted	254	32	11	7	1
4	Precision (row#2/row#1, %)	5	37	68	78	97
5	Recall (row #2/30 actual, %)	40	63	77	83	93

As expected, the citation distribution is skewed strongly towards many papers with few citations. We found that application of a lower citation cutoff of at least 5 citations within 6 months of publication resulted in elimination of more than 98.5% of papers but left 59% of breakthrough candidates (30 out of 51). A citation cutoff of 10 citations within 12 months dropped 97.7% of all papers but increased the number of BPs retained to 94% (48 out of 51).

As an example, we present results for the "5 citations within 6 months cutoff" that comprise 794 papers set in Tables 2 and 3. The quality of prediction versus initial detection time window was assessed by calculating precision (the ratio of actual BPs detected to the total number of predicted BPs) and recall (the ratio of actual BPs detected to the total number of BPs in data set).

In general, we found that the linear model tended to underestimate actual citations while the non-linear model tended to overestimate actual citation values. The linear model better predicts Type B and C citation patterns, or about 75% of all papers in the set. To demonstrate this statement, we analyzed top 200 ranked papers with cumulative citation count > 96. The histograms in Fig. 6 show the differences between predicted and actual cumulative count for linear (A1) and non-linear (B1) models are shown. Continuous curves (black color for linear model and red color for non-linear model) represent the best fit by the Student's *t*-distribution.

4. Conclusions

We have developed methods that combine curve-fitting and thresholding strategies for early detection of candidate breakthrough papers. We have shown the efficacy of this method empirically by testing precision and recall on known breakthrough papers and on detection of highly cited papers by topic area and across years. Our method is scalable to larger datasets, and is tunable by threshold and initial citation cutoff. A fraction of correctly predicted BPs increases with increase of the initial time window. For example, one year following publication, approximately 60% of BPs will be predicted correctly with 40% recall rate. Further experiments are needed to test and incorporate additional dimensions, such as interdisciplinarity of cited or citing articles, geographical diversity, and citations prestige to further refine the detection and qualification of candidate breakthrough papers. Even in this initial stage, our findings can be used to inform portfolio management practices. Application of an early detection indicator can be used to flag emerging research areas and stimulate attention through workshops, new funding, or collaborative research efforts.

Acknowledgments

We acknowledge the funding from the US National Institutes of Health contract # HHSN272200900231U which supported an initial phase of this study. We thank Y. Seger for the editorial assistance and Travis M. Hauguel for discussions.

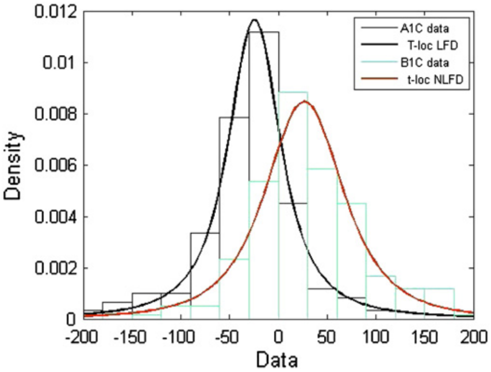

Fig. 6. Distribution of deviations of the predicted values from exact citation counts for 200 top ranked papers (see text). X-axis shows differences between predicted values and actual citations counts. Bins are normalized frequencies for linear (blue) and non-linear (green) models. Data fitted by the Student's *t*-distribution. Linear model fit (black) has a mean = −24.5, variance = 29.7, and ν (degrees of freedom) = 1.7. Non-linear model fit (red) has a mean = 26, variance = 42.8, and ν = 2.5.

References

[1] R. Rousseau, L. Egghe, Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science, Elsevier, Amsterdam, 1990.
[2] E. Garfield, Science 178 (1972) 471.
[3] H.G. Small, J. Am. Soc. Inf. Sci. 24 (1973) 265.
[4] D.J. De Solla Price, Science 149 (1965) 510.
[5] D.J. De Solla Price, J. Am. Soc. Inf. Sci. 27 (5-6) (1976) 292.
[6] R.K. Merton, Science 159 (3810) (1968) 56.
[7] A.-L. Barábasi, R. Albert, Science 286 (1999) 509.
[8] D. Pendlebury, The 1989 Nobel Prize in medicine: 20 who deserve it, Scientist 3 (19) (October 2 1989) 14.
[9] Eugene Garfield, Who Will win Nobel Prize in Economics? Curr. Contents 11 (March 12 1990) 3.
[10] Eugene Garfield, Alfred Welljams-Dorof, On Nobel Class: a citation perspectives on high impact research authors. (2) Theor. Med. 13 (1992) 117.
[11] www.Sciencewatch.com.
[12] E.S. Aversa, Scientometrics 7 (1985) 383.
[13] A.F.J. van Raan, Scientometrics 47 (2000) 347.
[14] S. Redner, Citation statistics from 110 years of physical review, Phys. Today 58 (2005) 49.
[15] Chaomei Chen, et al., Towards an explanatory and computational theory of scientific discovery, J. Informetr. 3 (July 3 2009) 191.
[16] Alan L. Porter, et al., Measuring researcher interdisciplinarity. (1) Scientometrics 72 (2007) 117.
[17] I. Rafols, M. Meyer, Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. (2) Scientometrics 82 (2010) 263.
[18] L. Leydesdorff, I. Rafols, Indicators of the interdisciplinarity of journals: diversity, centrality, and citations. (1) J. Informetr. 5 (2011) 87.
[19] S.M. Hammond, et al., An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells, Nature 404 (6775) (2000) 293.
[20] M.E. Domeier, et al., A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*, Science 289 (5486) (2000) 1928.
[21] R.F. Ketting, R.H. Plasterk, A genetic link between co-suppression and RNA interference in *C. elegans*, Nature 404 (6775) (2000) 296.
[22] N.J. Caplen, et al., dsRNA-mediated gene silencing in cultured Drosophila cells: a tissue culture model for the analysis of RNA interference, Gene 252 (2000) 95.
[23] M. Lagos-Quintana, et al., Identification of novel genes coding for small expressed RNAs, Science 294 (5543) (2001) 853.
[24] A. Fire, et al., Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*, Nature 391 (6669) (1998) 806.
[25] R.J. Distel, et al., Nucleoprotein complexes that regulate gene-expression in adipocyte differentiation — direct participation of c-fos, Cell 49 (6) (1987) 835.
[26] M.G. McHeyzer-Williams, M.M. Davis, Antigen-specific development of primary and memory t-cells in-vivo, Science 268 (5207) (1995) 106.
[27] H. Riezman, L. Hicke, Ubiquitination of a yeast plasma membrane receptor signals its ligand-stimulated endocytosis, Cell 84 (2) (1996) 277.
[28] A. Nussenzweig, et al., Requirement for Ku80 in growth and immunoglobulin V(D)J recombination, Nature 382 (6591) (1996) 551.
[29] J.D. Altman, et al., Phenotypic analysis of antigen-specific T lymphocytes, Science 274 (5284) (1996) 94.
[30] W. Glänzel, et al., Scientometrics 58 (3) (2003) 571.

Ilya V. Ponomarev, PhD, is a senior scientific analyst at Thomson Reuters, Custom Analytics and Engineered Solutions, based in Rockville, MD. He received a PhD in Theoretical Physics from the University of New South Wales, Australia. His research interests are in the areas of scientometrics, data mining, clustering algorithms, statistics, research and technology management, and science policy.

Duane E. Williams, PhD, is a senior scientific analyst at Thomson Reuters, Custom Analytics and Engineered Solutions, based in Rockville, MD. He received a PhD in Quantum Chemistry from the University of Florida. His current work focuses on the development and refinement of new metrics to quantify and characterize the impact of research funding on public health outcomes.

Charles J. Hackett, PhD, is the Deputy Director of the Division of Allergy, Immunology, and Transplantation (DAIT) of the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health. He received a PhD from Wayne State University and completed postgraduate training at the UK National Institute for Medical Research. He has served on the faculty of the Wistar Institute with an adjunct appointment at the University of Pennsylvania School Of Medicine, and was director of Cellular Immunology at ImmuLogic Pharmaceutical Corporation. His research focused on T cell responses to influenza virus and viral antigen presentation.

Joshua D. Schnell, PhD, is the Director of Analytics at Thomson Reuters, Custom Analytics and Engineered Solutions, based in Rockville, MD. He received a PhD in Biochemistry from Northwestern University and was a National Academies fellow and served as Assistant Chair of the Department of Biochemistry, Molecular Biology, and Cellular Biology at Northwestern University before joining Thomson Reuters. His areas of expertise include research training, program evaluation, and development of research indicators.

Laurel L. Haak, PhD was the Chief Science Officer at Thomson Reuters, Custom Analytics and Engineered Solutions, based in Rockville, MD. She received a PhD in Neurosciences from Stanford University School of Medicine and performed postdoctoral research at the National Institutes of Health. She has served as an Editor at the American Association for the Advancement of Science (AAAS) and Program Manager at the US National Academies. Her areas of expertise include science policy, science workforce dynamics, and research program evaluation. She is currently the Executive Director of ORCID.