Ponomarev, Ilya; Williams, Duane; Lawton, Brian; Cross, Di H.; Seger, Yvette; Schnell, Joshua; Haak, Laurel: *Breakthrough Paper Indicator: early detection and measurement of ground-breaking research.* In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 295-304. ISBN 978-80-86742-33-5. Available from: www.eurocris.org.

Breakthrough Paper Indicator: early detection and measurement of ground-breaking research

Ilya V. Ponomarev, Duane E. Williams, Brian K. Lawton, Di H. Cross,
Yvette Seger, Joshua Schnell, Laurel L. Haak
Thomson Reuters, Rockville, MD, USA

Summary

Breakthrough scientific discoveries are rare occurrences that have a huge impact on technological progress, however, the full potential of such breakthroughs are recognized retrospectively and on a time scale of decades. Thus, rapid detection and recognition of potential breakthroughs is a core goal of research management and science policy. Here we describe methods for early detection of candidate breakthrough publications based on the dynamics of publication citations and citation quality. We used a quantitative approach to identify typical citation patterns of highly cited papers. Based on these analyses, we propose two forecasting models that allow us to select groups of breakthrough paper candidates that exceed high citation thresholds five years post-publication. In the second phase, diversity in the subject categories of papers was studied as a possible candidate measure to improve ranking of breakthrough paper candidates. Models were validated using statistical methods to derive confidence levels.

1 Introduction

Scientific knowledge typically evolves through incremental steps; however abrupt discoveries have the potential to reshape an entire field of study. Identification and measurement of the impact of rare research breakthroughs are vital from both a management and policy standpoint. By detecting ground-breaking research as early as possible, more time is gained to support new emerging technologies through workshops, new funding, or collaborative research efforts.

The quantitative study of ground-breaking research publications has a long history, much of which has been based 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 the term "cumulative advantage" to describe the mechanism that causes a persistently higher citation rate. In the framework of network models, this mechanism is now known as preferential attachment [5]. In the 1980s, Pendlebury [6] and Garfield [7] performed analyses of total citation counts of eminent researchers and showed that most-cited author ranking effectively identified Nobel Prize winners [8]. This approach is used currently in the Essential Science Indicators^{SM} (ESI) product from Thomson Reuters [9]. More recently, Redner [10] 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,

Ponomarev, Ilya; Williams, Duane; Lawton, Brian; Cross, Di H.; Seger, Yvette; Schnell, Joshua; Haak, Laurel: *Breakthrough Paper Indicator: early detection and measurement of ground-breaking research.* In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 295-304. ISBN 978-80-86742-33-5. Available from: www.eurocris.org.

highly correlated bursts of citations, and down-turns in research activity. Chen et al. [11] introduced the explanatory and computational theory of transformative discovery based on a network approach to scientific knowledge diffusion.

In this manuscript, we report progress on new developments in the breakthrough paper indicator. Our earlier studies [26] resulted in the development of forecasting models, which describe and 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. These models allow us to select a group of breakthrough paper candidates with predicted high citation counts. It is worth noting that, much like research itself, identification of influential discoveries is a multidimensional process and should involve metrics beyond simple cumulative citation counts. Examples include ranking citations by geographic region, by interdisciplinary features [12-14], by prestige diversity (e.g., fraction of citations published in high impact journals), 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 next step in the development of such a multidimensional breakthrough paper indicator. Here we describe our forecasting models, as well as analyze the quality of citations by studying subject category diversity of citing papers.

2 Method and Analysis

We use ranked citation counts and monthly citation rates as proxies for scientific impact. The following steps briefly describe our approach:

1. A small set of known BPs in a particular research field (biochemistry and molecular biology) was used as a test for indicator development.
2. For each BP, a statistically large set of similar publications was identified.
3. Each paper within set was then ranked by cumulative citations, and citation breakthrough thresholds established and justified.
4. A typical pattern of time-dependent citation behavior of highly cited papers was identified.
5. The theoretical model best describing this citation behavior with a high level of statistical confidence was selected, using knowledge of early citation patterns (first 6-24 months following publication) to predict later citation dynamics.
6. For those candidates whose predicted citations exceeded the breakthrough threshold after 60 months, we studied the diversity of subject categories for their references and citations to further refine ranking.

2.1 Citation data

For this study, we used citation data sets derived from Thomson Reuters Web of Science® (WoS) that were extracted in June 2011. An initial set of 11 known breakthrough publications in biochemistry & molecular biology was compiled with input from 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 (NIH) [15-25]. In our analysis, this set is referred to as "DAIT BPs" (see Table 1).

For each DAIT paper, we identified a set of similar papers. A paper was considered similar to a DAIT breakthrough publication if it was published in the same journal during the same calendar year. This allowed us to generate a data corpus capable of supporting statistical significance testing. General information about DAIT BPs and corresponding similar paper sets is provided in Table 1.

Table 1: Citation and diversity statistics of DAIT BPs. '1st Author Name' – first author names in reference set [15-25]; '# Sim Pubs' – number of similar publications in the comparison set; 'Cites 2011' – total number of citations received as of June 2011; 'Cites 60M' – number of citations after 60 months following publication; 'Citations Rank' – citation rank of the paper in the set (#1 corresponds to the highest cited paper); 'Diversity rank – rank of paper based on diversity index of citations (see Section 2.5); 'Cumulative Rank' – cumulative paper rank .

1st Author Name	Journal	Pub Year	# Sim Pubs	Cites 2011	Cites 60M	Citations Rank	Diversity Rank	Cumulative Rank
Hammond	NATURE	2000	5,106	1,317	606	39	367	70
Ketting	NATURE	2000	5,106	116	88	1,225	1,139	793
Domeier	SCIENCE	2000	5,106	106	75	1,501	2,099	1,281
Caplen	GENE	2000	33,210	131	92	1,857	2,167	810
Lagos-Quintana	SCIENCE	2001	5,136	1,267	491	68	545	111
Fire	NATURE	1998	5,300	4,856	998	13	966	227
Distel	CELL	1987	18,567	498	324	60	293	60
McHeyzer-Williams	SCIENCE	1995	4,899	351	166	480	3,891	2,096
Hicke	CELL	1996	29,664	513	253	232	6,013	1,655
Nussenzweig	NATURE	1996	5,153	424	203	314	1,116	427
Altman	SCIENCE	1996	5,153	2,236	701	30	80	541

Both WoS subject categories (S=263 as of June 2011) and ESI standard categories (S=22) were used to classify research field. To estimate breakthrough citation thresholds for different research fields, we evaluated annual data sets of research articles published from 1995-2005 in 22 ESI subject categories that acknowledged NIH funding support (referred to as "MEDLINE Sets"). Findings were validated using the 2005 data set (total 375,372 items). Citation data were analyzed in one-month intervals.

2.2 Citation breakthrough threshold

Our initial retrospective analysis was performed using the 2005 MEDLINE data set. We established a 5 year time window following publication date for calculation of cumulative citation rankings, with the assumption that this is a sufficient time interval for seminal papers to be recognized by the scientific community. A candidate breakthrough was defined as a paper that exceeded a certain threshold of cumulative citation count 5 years after publication. A comparison of cumulative citation distributions between the DAIT BP and 2005 MEDLINE sets is shown in Figure 1.

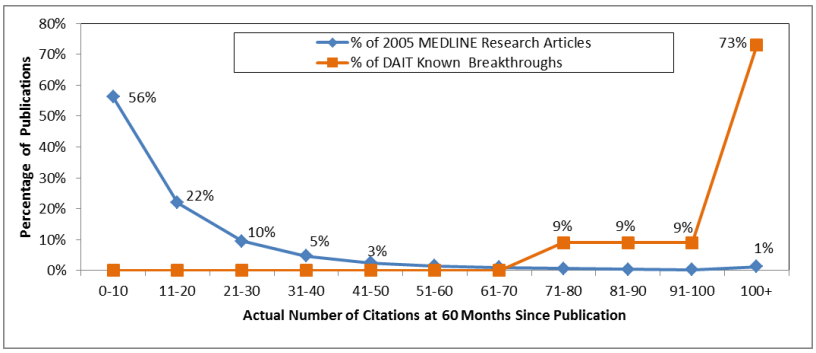

Figure 1: Distribution of cumulative number of citations after 5 years for two data sets. MEDLINE set in the 2005 data set (blue, 375K articles) shows typical highly skewed power law behavior with only about 1% papers being cited more than 100 times. In comparison, all DAIT BPs (orange) have been cited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

Thus, a key question for this analysis is how to determine 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 this does not account for variation in citation behavior between fields. This is clearly shown in Figure 2, in which a threshold of 100 citations would result in 1,858 BP candidates in Clinical Medicine in 2005 alone, and more than 500 papers in 4 other subject categories. Such a high rate of BPs does not correspond to the intuitive expectation regarding the rarity of breakthrough discoveries. Conversely, the use of an arbitrary high threshold would leave some categories such as Space Science or Economics with no breakthrough candidates. This variance in the number of citations is dependent on total publication volume and the size of particular scientific community, which can vary substantially by field of research (see right column in Figure 2).

We determined that a percentile approach was more effective than a strict numerical cutoff for establishing the citation threshold value. In general, the top percentage of publications within a set should be selected as candidates, while recognizing the need to balance selectivity with inclusiveness. Once the candidate groups are selected, other metrics may be applied to filter the results further. We determined that for topical sets containing more than 20,000 publications, a 0.1% cutoff is optimal (Figure 2). For smaller sets, it may be desirable to increase the cutoff to 1% or more.

To test whether the citation threshold changes over time, we analyzed annual data sets derived from MEDLINE for each of the years 1995-2004. While some of subject categories have monotonic dependencies or strong fluctuations (e.g., material science, chemistry), we found that re-

search fields related to medicine and biology have a stable citation threshold. This is a very important finding since it allows us to project the current average value of the citation threshold in future years.

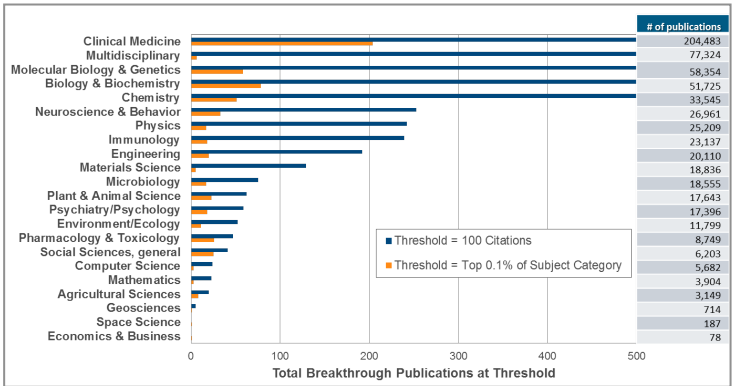

Figure 2: BPs citation thresholds and corresponding number of papers, by subject category for numeric (blue) and percentile (orange) threshold selection.

2.3 Identifying citations patterns

The next step in our detection method was identification of citation patterns of top-cited publications using monthly citation counts and rates. In Figure 3, cumulative citation counts (left panel) and their derivatives – a monthly citation rate (right panel) – are shown. A typical citation pattern has an initial period of slow citation growth lasting from 5 to 20 months (monthly rate is proportional to t^{α} with $1 \le \alpha \le 2$). After this initial slow growth phase, the citation rates accelerate until they reach saturation plateaus, after which they decrease (memory or aging effect). While the citation counts for the majority of top ranked publications will follow this scenario, the time transitions between these phases vary substantially from paper to paper. When analyzing the top cited papers in several datasets, we found that approximately 25% of them are still at the first stage of growth after 5 years (Type A, right panel of Figure 3), 50% are in saturation (Type B) and 25% have started aging (Type C).

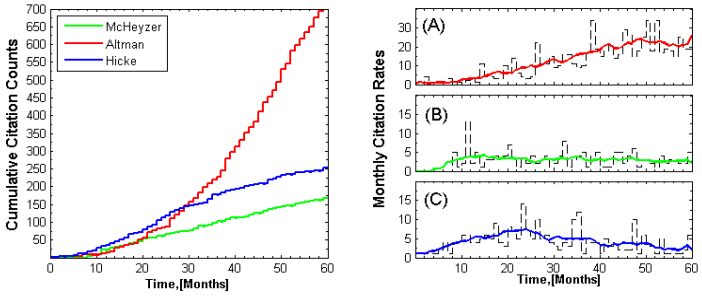

Figure 3: Three typical citation count patterns. Due to strong fluctuations, monthly citation rates were smoothed by applying a moving average window with 5 month span. Please note different upper limits for y-axis for monthly citation rates.

2.4 Forecasting Models

Based on these observations, we developed a forecasting model that allows us to predict most citation trajectories with a minimum number of fitting parameters. Our approach was as follows: using data for the first n months (n=6, 12, or 24), we curve-fit citation data and extrapolated results to time t=5 years and predicted whether or not the paper will be above the citation threshold for a given subject category (has breakthrough potential). We found that for predictions to have a strong statistical confidence level the paper under consideration has to receive enough citations during the initial fitting period, otherwise the scarcity of data introduces large random fluctuations. Therefore, we discarded publications which did not have at least 5 citations at 6 months from publication date for our analysis. In doing so, we risked omission of a group of papers which became highly cited later (aka "sleeping beauties" or "late bloomers"), but this group is always difficult to identify at early stage.

In our method, we chose linear model (LM) f_1 and non-linear model (NLM) f_2 theoretical curves to fit the citation behavior (Figure 4):

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

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

We used the 2005 MEDLINE BPs for the Chemistry subject category as a test set. This set comprised 169 journals and 51,575 publications. Of these, 51 papers are above the proposed 0.1% citation breakthrough threshold (262 citations after 5 years). First, we experimented with optimization of the initial detection time window. We found that 6 months provided an appropriate metric for eliminating a larger percentage of non-breakthrough papers while retaining approximately 60% of breakthroughs. A window of 12 to 24 months provides sufficient time to identify key citation patterns while retaining approximately 90% of BPs. As expected, the citation distribution is skewed strongly by many papers with few citations. Application of the lower citation cutoff of 5 citations after 6 months allowed us to eliminate more than 98% of papers (including 21 (40%) breakthrough candidates). Following this cutoff, 794 papers were left in the set. Models were assessed using 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 and 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.

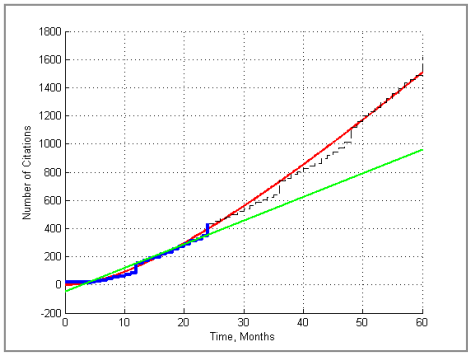

Figure 4: Citation curve of actual paper (dashed line), linear (green) and nonlinear (red) fitting curves. Fitting procedure used only citation data for the first several months (shown in blue). Data were extrapolated to predict future citations and compared with actual citation curve.

2.5 Measuring diversity of subject categories of citations

Table 1 indicates that some known BPs have low cumulative citations ranks. For example, "Ketting", "Domeier", and "Caplen" papers were outperformed by more than one thousand similar papers. This confirms the known fact that the number of citations is merely an approximate proxy for scientific impact, and supplementary information is often necessary to distinguish true path-finding discovery from incremental knowledge generation. We studied publication interdisciplinarity as one such complimentary information parameter. The reduction of disciplinary barriers between scientific disciplines through interdisciplinary research is seen as critical to expediting progress in research and development [12]. Interdisciplinary research (IDR) is defined as research by teams or individuals that integrate perspectives, concepts, or theories; tools, or techniques; and/or information, or data from two or more bodies of specialized knowledge or research practice [14]. Its purpose is to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single field of research practice. Therefore, we sought to determine whether breakthrough papers were more interdisciplinary in their nature or not.

We selected several standard measures of diversity (richness, exponential of Shannon entropy, and inverse Simpson concentration) as well as the Rao-Stirling-Porter interdisciplinarity index. Required elements for calculating standard diversities are:

1. Reference and citation lists for publications of interest.
2. A categorization of those references/cites into disciplinary categories (journal subject categories)

The interdisciplinarity index also requires a third element: a measure of similarity between discipline categories (cosine of journal subject category pairs) - taken from a baseline or comparison set of publications which may or may not be outside the body of literature of interest for calculating the integration index. Roughly speaking, all indices calculate the effective number of different subject categories. In addition, the interdisciplinarity index also tries to account for the degree to which the subject categories are (or are not) different from one another by weighting the inclusion of two very similar fields less than the inclusion of two very different fields.

If p_i, p_j are proportions of i^{th} and j^{th} subject categories from all set of SCs ($i, j=1,...,S$) in references or citations list for the chosen publication, then the diversity index of order q is defined by formula:

$$D_q = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}}, q \neq 1$$

$$D_1 = \exp\left[-\sum_{i=1}^{S} p_i \log\left(p_i \right) \right], q = 1.$$

The diversity of order zero is known as richness (simply a number of different SCs in a set), the diversity of order 1 is the exponent of Shannon's entropy, and the diversity of order 2 is the inverse Simpson concentration. The formula for Rao-Stirling-Porter interdisciplinarity is

$$I = 1 - \sum_{i \leq j}^{S} s_{ij} p_i p_j$$

This measure is also known in the literature as "quadratic entropy" because, unlike traditional measures of diversity, the probability distributions (p_i and p_j) of the references/cites are multiplied by the distance (cosine similarity) in the (reference/citation) network among them (s_{ij}). The latter factor represents the journal ecology: a journal which is interdisciplinary within a specific domain may not reach beyond the confines of this domain. Other journals can be highly specialized, yet combine citations from or to different domains. We calculated diversity measures for all DAIT sets. Table 2 shows an example of calculations for the "Distel" paper.

Table 2: Diversity/interdisciplinarity measures for "Distel" paper (first 6 columns) and all similar papers (last two columns) in corresponding DAIT set

Type	# of pubs	D_0	$D_{1/2}$	D_1	D_2	I (RSP)	# of pubs in similar sets	D_2^{sim}
Refs	34	4	3.2	2.9	2.6	0.36	526208	6.1
Cites	324	31	18.8	12.3	7.5	0.59	553336	8.2

We found that all diversity measures are highly correlated between each other. However, we did not find any correlations between citation rank and diversity ranks (compare the third to last and second to last columns in Table 1) within the DAIT sets. Namely, papers with high citation counts can have a very low number of journal subject categories or narrow area of focus, while papers with a low citation rank can be highly "diverse". One explanations for this result is that similar publications from multidisciplinary journals such as "Nature" and "Science" are not interdisciplinary themselves; rather they belong to distinct research fields (medicine, biochemistry, physics, history) which have different reference and citation patterns. Therefore, additional subject category assignment at article level is required for these journals. On the other hand, we observed that the citation diversity (including RSP index) correlates very weakly with cumulative number of citations, even for a single, well-defined subject category.

For a true measure of interdisciplinarity, we propose that the difference between citation and reference patterns has to be measured together rather than treated separately. An illustration of this is shown in Figure 5 in which you can see that the distribution of citations and references for all similar articles in the set have the same pattern (it is also confirmed by small difference be-

tween $D_2^{\text{sim cites}}$ and $D_2^{\text{sim refs}}$). However, the pattern of references and citations for the "Distel" paper itself is quite different; the paper referenced essentially 3 subject categories while it received citations from a broader spectrum of research fields ($D_2^{\text{cites}} - D_2^{\text{refs}} = 4.9$). We believe this observation may be captured using the diversity measure based on the Kolmogorov-Smirnov distance [27].

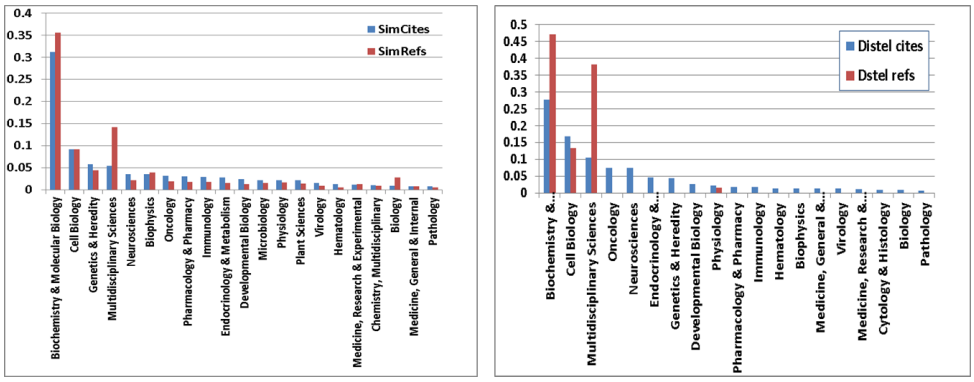

Figure 5: Distribution of normalized frequencies (p_i) for citations and references for all similar publications in "Distel" similarity set [left], for citations and references of "Distel" paper only[Right].

3 Conclusions

In this paper, we report the development of methods that combine curve-fitting and thresholding strategies to allow for the early detection of candidate breakthrough papers. We have empirically shown the efficacy of method to detect highly cited papers by topic area and across years by testing precision and recall using a set of known breakthrough papers. Our method is scalable to larger datasets, and is tunable by threshold. We also studied different diversity measures applied to references and citations. Further experimentation is required to incorporate and test additional dimensions, such as geographical diversity, and citation 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.

References

1. Rousseau, L.; Egghe, R. (1990): *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science.* Amsterdam : Elsevier
2. Garfield, E. (1972): *Science,* Vol. 178, p. 471.
3. Small, H. G. (1973): *J. Amer. Soc. Inform. Sci.,* Vol. 24, p. 265.
4. Price, D. J. De Solla (1965): *Science,* Vol. 149, p. 510.

5. Barábasi, A.-L. and Albert, R. (1999)*: Science,* Vol. 286, p. 509.
6. Pendlebury, D. (October 2, 1989):*Scientist*, Vol. 3 (19), p. 14.
7. Garfield, E. (March 12, 1990: *Current Contents,* Vol. 11, p. 3.
8. Garfield, E.; Welljams-Dorof, A. (1992): *Theor. Med*, Vol. 13, p. 117.
9. www.Sciencewatch.com.
10. Redner, S. (2005): *Physics Today*, Vol. 58, p. 49.
11. Chen, Chaomei, et al., (2009): *Journal of Informetrics*, Vol. 3, p. 191.
12. Porter, A. L; et al. (2007): *Scientometrics*, Vol. 72, p. 117.
13. Rafols, I. ; Meyer, M. (2010): *Scientometrics*, Vol. 82, p. 263.
14. Leydesdorff, L.; Rafols, I. (2011): *Journal of Informetrics,* Vol. 5, p. 87.
15. Hammond, S M; et al. (2000): *Nature*, Vol. 404 (6775), p. 293.
16. Domeier, M. E.; et al. (2000) *Science,* Vol. 289 (5486), p. 1928.
17. Ketting, R. F. ; Plasterk, R H. (2000): *Nature,* Vol. 404 (6775), p. 296.
18. Caplen, N J; et al. (2000): *Gene,* Vol. 252, p. 95.
19. Lagos-Quintana, M.; et al. (2001): *Science,* Vol. 294 (5543), p. 853.
20. Fire, A.; et al. (1998): *Nature,* Vol. 391 (6669), p. 806.
21. Distel, R. J.; et al. (1987): *Cell,* Vol. 49 (6), p. 835.
22. McHeyzer-Williams, M. G.; Davis, M. M. (1995): *Science,* Vol. 268 (5207), p. 106.
23. Riezman, H.; Hicke, L. (1996): *Cell,* Vol. 84 (2), p. 277.
24. Nussenzweig, A.; et al. (1996): *Nature,* Vol. 382 (6591), p. 551.
25. Altman, J. D.; et al. (1996): *Science*, Vol. 274 (5284), p. 94.
26. Ponomarev I. V.; et al. (2012): to be published in *Technological Forecasting and Social Change,* Vol. 79 (12).
27. W.J.Conover (1999): *Practical Nonparametric Statistics*, 3rd ed., NYork: John Willey & Sons

Contact Information

Ilya V. Ponomarev

Thomson Reuters
1455 Research Blvd
2nd Floor
Rockville
MD 20850
USA

ilya.ponomarev@thomsonreuters.com

