SIGNIFICANCE ROYAL STATISTICAL SOCIETY ASA AMERICAN STATISTICAL ASSOCIATION Statistical Society of Australia

Volume 3, Issue 4
December 2006
Pages 179-181

Do you sincerely want to be cited? Or: read before you cite

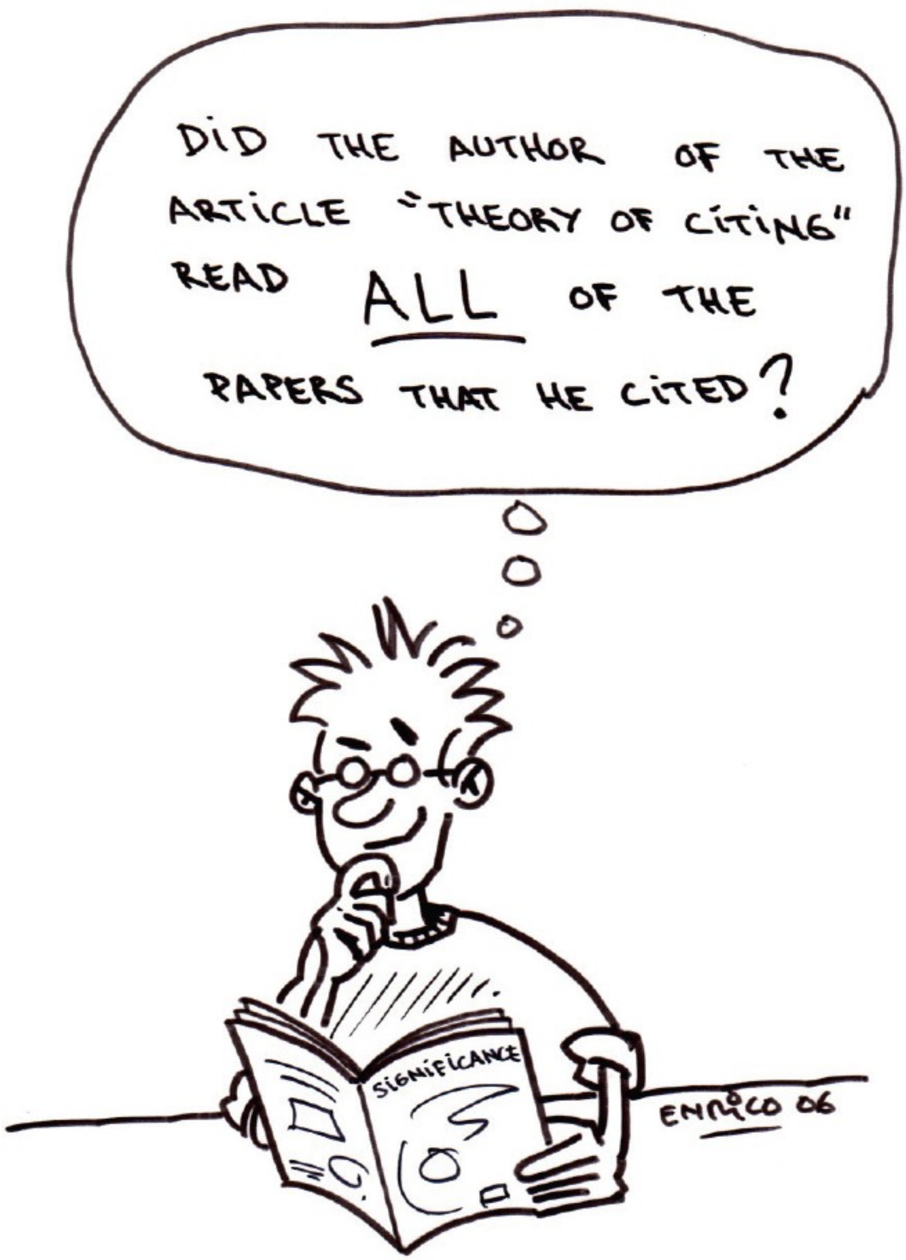
Mikhail Simkin, Vwany Roychowdhury

First published: 28 November 2006 |
https://doi.org/10.1111/j.1740-9713.2006.00202.x | Citations: 21

Abstract

Do you sincerely want to be cited? Prestige depends on the number of times your academic paper gets cited. But that need not be a measure of how good it is, nor even of how many times it is actually read. **Mikhail Simkin** and **Vwani Roychowdhury** explain their theory of the unread citation.

Many psychological tests have the so-called lie-scale. A small but sufficient number of questions that admit only one true answer, such as "Do you *always* reply to letters immediately after reading them?", are inserted among others that are central to the particular test. A wrong reply for such a question adds a point on the lie-scale, and when the lie-score is high, the overall test results are discarded as unreliable. Perhaps for a scientist the best candidate for such a lie-scale is the question "Do you read *all* of the papers that you cite?".



Comparative studies of the popularity of scientific papers have been a subject of much interest[1-4], but the scope has been limited to citation-counting. We discovered a method of estimating what percentage of people who cite a paper have actually read it[5]. Remarkably, this can be achieved

[Preface: 'Retrospective read paper'](#)

David Firth

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

[Do We Really Want Productivity?](#)

H. D. Dickinson

Journal of the Royal Statistical Society: Series A (General)

[A mathematical theory of citing](#)

Mikhail V. Simkin, Vwani P. Roychowdhury

without any testing of the scientists, but solely on the basis of the information available in the ISI citation database (available from www.isiwebofknowledge.com).

Freud[6] discovered that the application of his technique of psychoanalysis to slips in speech and writing could reveal a lot of hidden information about human psychology. Similarly, we find that the application of statistical analysis to misprints in scientific citations can give an insight into the process of scientific writing. As in the Freudian case, the truth revealed is embarrassing. For example, an interesting statistic revealed in our study is that a lot of misprints are identical. The probability of repeating someone else's misprint accidentally is small. One concludes that repeat misprints are most likely to occur when copying from a reference list used in another paper.

Our initial report[5] led to a lively discussion (see http://science.slashdot.org/article.pl?sid=02/12/14/0115243&mode=thread&tid=134 on whether copying citations is tantamount to not reading the original paper. Alternative explanations are worth exploring, although such hypotheses should be supported by data and not by anecdotal claims. It is indeed most natural to assume that a copying citer has also failed to read the paper in question (albeit this cannot be rigorously proved). *Entities must not be multiplied beyond necessity*. Having thus shaved the critique with Occam's razor, we will proceed to use the term non-reader to describe a citer who copies.

As misprints in citations are not too frequent, only celebrated papers provide enough statistics to work with. Let us have a look at the distribution of misprints in citations to one renowned paper[7], which accumulated 4300 citations (though the misprint distribution for a dozen of other studied papers is very similar[8]). Of these citations, 196 contain misprints, out of which only 45 are distinct. The most popular misprint, in a page number, appeared 78 times.

> **Statistical analysis of scientific citations reveals embarrassing truths**

As a preliminary attempt, one can estimate the ratio of the number of readers to the number of citers, *R*, as the ratio of the number of **distinct** misprints, *D*, to the **total number** of misprints, *T*. Clearly, among *T* citers, *T* – *D* copied, because they repeated someone else's misprint. For the *D* others, with the information at hand, we have no evidence that they did not read, so according to the presumed innocent principle, we assume that they did. Then in our sample, we have *D* readers and *T* citers, which leads to:

$$R \approx D/T \tag{1}$$

Substituting *D* = 45 and *T* = 196 into equation 1, *R* ã *D/T*, we obtain *R* ã 0.23. This estimate would be correct if the people who introduced original misprints had always read the original paper. It is more reasonable to assume that the probability of introducing a new misprint in a citation does not depend on whether the author has read the original paper. Then, if the fraction of read citations is *R*, the number of readers in our sample is *RD*, and the ratio of the number of readers to the number of citers in the sample is *RD/T*. What happens to our estimate in equation 1? It is correct, just the sample is not representative: the fraction of read citations among the citations containing misprints is less than in the general citation population.

Can we still determine R from our data? Yes. From the misprint statistics we can determine the average number of times, n_p, a typical misprint propagates:

$$n_p = \frac{T - D}{D} \tag{2}$$

The number of times a misprint had propagated is the number of times the citation was copied from either the paper that introduced the original misprint, or from one of the subsequent papers that copied (or copied from copied etc.) from it. A misprinted citation should be no different from a correct citation as far as copying is concerned. This means that a selected-at-random citation, on average, is copied (including copied from copied etc.) n_p times. The read citations are no different from unread citations as far as copying goes. Therefore, every read citation, on average, was copied n_p times. The fraction of read citations is thus:

$$R = \frac{1}{1 + n_p} \tag{3}$$

After substituting equation 2 into equation 3, we recover equation 1.

Note, however, that the average number of times a misprint propagates is not equal to the number of times the citation was copied, but to the number of times it was copied *correctly*. Let us denote the average number of citations copied (including copied from copied etc.) from a particular citation as n_c, which can be determined from n_p in the following way. The n_c consists of two parts: n_p (the correctly copied citations) and misprinted citations. If the probability of making a misprint is M and the number of correctly copied citations is n_p, then the total number of copied citations is

$$\frac{n_p M}{1 - M}$$

and the number of misprinted citations is

$$\frac{n_p}{1 - M}$$

As each misprinted citation was itself copied n_c times, we have the following self-consistency equation for n_c:

$$n_c = n_p + n_p \times \frac{M}{1 - M} + (1 + N_c) \tag{4}$$

Equation 4 has the solution

$$n_c = \frac{n_p}{1 - M - n_p \times M} \tag{5}$$

After substituting equation 2 into equation 5 we get:

$$n_c = \frac{T - D}{D - MT}$$

(6)

From this we get:

$$R = \frac{1}{1 + n_c} = \frac{D}{T} \times \frac{1 - (MT)/D}{1 - M}$$

(7)

The probability of making a misprint can be estimated as

$$M = \frac{D}{N}$$

where *N* is the total number of citations. After substituting this into equation 7 we get:

$$R = \frac{D}{T} \times \frac{N - T}{N - D}$$

(8)

Substituting *D* = 45, *T* = 196, and *N* = 4300 into equation 8, we get *R* ã 0.22, which is very close to the initial estimate obtained using equation 1.

Copied citations create renowned papers

During the Manhattan project, the making of the nuclear bomb, Fermi asked General Groves, the head of the project, what would be the definition of a "great" general[9]. Groves replied that any general who had won five battles in a row might safely be called great. Fermi then asked how many generals are great. Groves said about three out of every hundred. Fermi conjectured that, considering that opposing forces for most battles are roughly equal in strength, the chance of winning one battle is 1/2 and the chance of winning five battles in a row is $1/2^5 = 1/32$. “So you are right General, about three out of every hundred. Mathematical probability, not genius.” The existence of military genius was also questioned on basic philosophical grounds by Tolstoy[10].

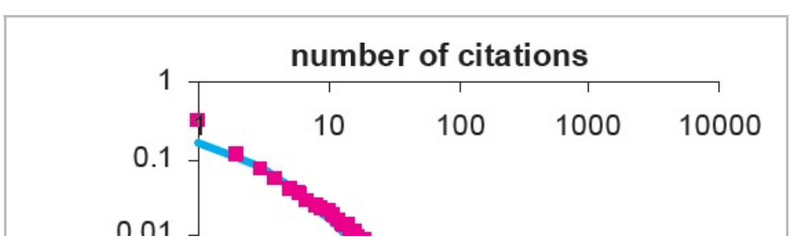
A commonly accepted measure of “greatness” for scientists is the number of citations to their papers[2]. For example, SPIRES, the high-energy physics literature database (http://www.slac.stanford.edu/spires/), divides papers into six categories according to the number of citations they receive. The top category, renowned, papers are those with 500 or more citations. Let us have a look at the citations to roughly 24 000 papers, published in *Physical Review D* between 1975 and 1994 (SPIRES data compiled by H. Galic, and made available by S. Redner (http://physics.bu.edu/~redner/projects/citation/). As of 1997 there where about 350 000 such citations: 15 per published paper on average. However, 44 papers were cited 500 times or more. Could this happen if all papers are created equal?

If they indeed are, then the chance of winning a citation is one in 24 000. What is the chance of winning 500 cites out of 350 000? The calculation is slightly more complex than in the militaristic case, but the answer is one in 10^{500}. In other words, it is zero. One is tempted to conclude that those 44 papers that achieved the impossible are indeed great.

A more careful analysis puts this conclusion in doubt. We just have shown that the majority of scientific citations are copied from the lists of references used in another papers. This way a paper that already was cited is likely to be cited again, and after it is cited again it is even more likely to be cited in the future. In other words, "unto every one that hath shall be given, and he shall have abundance[11]". This phenomenon is known as either the Matthew effect[12], cumulative advantage[13], or preferential attachment[14].

The effect of citation copying on the probability distribution of citations can be quantitatively understood within the framework of the model of random-citing scientists (RCS)[15, 16], which is as follows. When a scientist is writing a manuscript he picks up m random articles, cites them, and also copies some of their references, each with probability p.

This model was stimulated by the recursive literature search model[17] and can be solved using methods developed to deal with multiplicative stochastic processes[18]. These methods are too complicated to be described in a popular article so we will just state the results. A good agreement between the RCS model and actual citation data (see http://science.slashdot.org/article.pl?sid=02/12/14/0115243&mode=thread&tid=134) is achieved with the input parameters $m = 3$ and $p = 1/4$ (see Figure 1). Now what is the probability for an arbitrary paper to become renowned, i.e. receive more than 500 citations? A calculation shows that this probability is one in 600. This means that about 40 out of 24 000 papers should be renowned; ergo, mathematical probability, not genius.



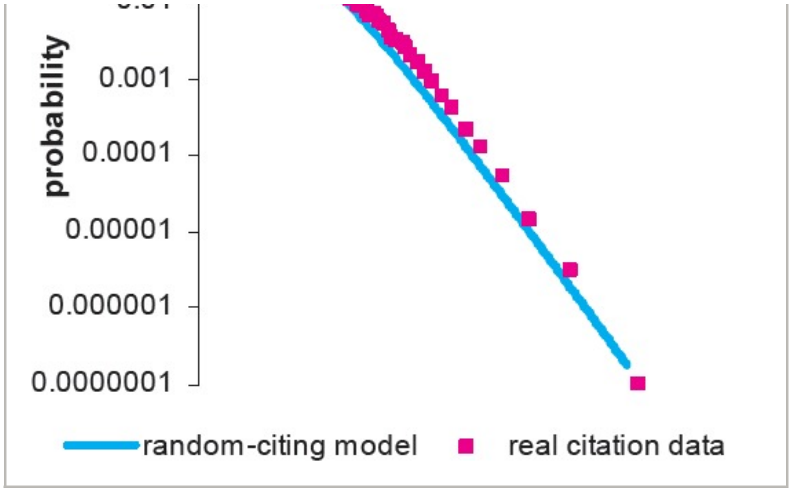
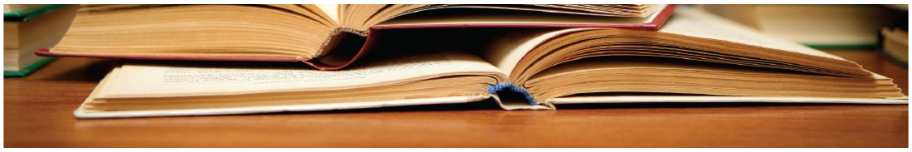


Figure 1 Open in figure viewer | PowerPoint

The outcome of the model of random citing compared to actual citation data. Mathematical probability rather than genius can explain why some papers are cited a lot more than others

In one incident[19] Napoleon (incidentally, he was the military commander whose genius was questioned in *War and Peace*[10]) said to Laplace "They tell me you have written this large book on the system of the universe, and have never even mentioned its Creator." The reply was "I have no need for this hypothesis." It is worthwhile to note that Laplace was not against God. He simply did not need to postulate his existence in order to explain existing astronomical data. Similarly, the present work is not blasphemy. Of course, in some spiritual sense, great scientists do exist. It is just that even if they did not exist, the citation data would look the same.



References

1 Price, D. de S. (1965) Networks of Scientific Papers. *Science*, **149**, 510.
Crossref | CAS | PubMed | Web of Science® | Google Scholar

2 Garfield, E. (1979) *Citation Indexing*. New York: John Wiley.
Google Scholar

3 Silagadze, Z. K. (1997) Citations and Zipf-Mandelbrot law. *Complex Systems*, **11**, 487.
Google Scholar

4 Redner, S. (1998) How popular is your paper? An empirical study of citation distribution. *European Physics Journal B*, **4**, 131.
Crossref | CAS | Web of Science® | Google Scholar

5 Simkin, M. V. and Roychowdhury, V. P. (2003) Read before you cite! *Complex Systems*, **14**, 269. (Available from http://arxiv.org/abs/cond-mat/0212043).
Google Scholar

6 Freud, S. (1901) Zur Psychopathologie des Alltagslebens.
Google Scholar

7 Kosterlitz, J. M. and Thouless, D. J. (1973) *Journal of Physics C*, **6**, 1181– 1203.
Crossref | CAS | PubMed | Web of Science® | Google Scholar

8 Simkin, M. V. and Roychowdhury, V. P. (2005) Stochastic

modeling of citation slips. *Scientometrics*, **62**, 367– 384. (Available from http://arxiv.org/abs/condmat/0401529 .).
Crossref | Web of Science® | Google Scholar

9 Deming, W. E. (1986) *Out of the crisis cambridge*: MIT.
Google Scholar

10 Tolstoy, L. (1869) *War and Peace*.
Google Scholar

11 *Gospel according to Matthew* **25**: 29.
Google Scholar

12 Merton, R. K. (1968) The Matthew Effect in Science. *Science*, **159**, 56. In fact, similar sayings appears in two other gospels: "For he that hath, to him shall be given…" [Mark 4:25], "…unto every one which hath shall be given…" [Luke 19:26] and belong to Jesus. Nonetheless the name "Matthew effect" has been repeated by thousands of people who do not read the Bible.
Crossref | CAS | PubMed | Web of Science® | Google Scholar

13 Price, D. de S. (1976) A general theory of bibliometric and other cumulative advantage processes. *Journal of American Society for Information Science*, **27**, 292.
Wiley Online Library | Web of Science® | Google Scholar

14 Barabasi, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509.
Crossref | CAS | PubMed | Web of Science® | Google Scholar

15 Simkin, M. V. and Roychowdhury, V. P. (2005) *Copied citations create renowned papers*? Annals of Improbable Research, Jan.-Feb., 24– 27. (Available from http://arxiv.org/abs/cond-mat/0305150 .).
Google Scholar

16 Simkin, M. V. and Roychowdhury, V. P. (0000) A

mathematical theory of citing. (Available from http://arxiv.org/abs/cond-mat/0504094 .).
Google Scholar

17 Simon, H. A. (1957) *Models of Man*. New York: Wiley.
Crossref | Google Scholar

18 Vazquez, A. (2001) Knowing a network by walking on it: emergence of scaling. *Europhysics Letters*, **54**, 430. (Available from http://arxiv.org/abs/cond-mat/0006132 .).
Crossref | CAS | Web of Science® | Google Scholar

19 De Morgan, A. (1915) *A budget of paradoxes*. Chicago: The Open Court Publishing Co. Vol. **2**, p. 1.
Google Scholar