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

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, 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. Remarkably, this can be achieved 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 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 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 paper7, which accumulated 4300 citations. Substituting equation 2 into equation 3, we get:

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

(8)

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

$$R = \frac{D}{T} (N - T) (N - D)$$  

(9)

Substituting $D = 45$, $T = 196$, and $N = 4300$ into equation 8, we get $R = 0.22$, 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”. 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 Tolstoy5. A commonly accepted measure of “greatness” for scientists is the number of citations to their papers. 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. Gaits, and made available by S. Redner (http://physics.bu.edu/~redner/projects/citation/)). As of 1997 there were 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. 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 $10^{36}$. In other words, it is zero. One is tempted to conclude that these 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, which is as follows. When a scientist is writing a manuscript he picks up random articles, cites some of them, and also copies some of their references, each with probability $p$. This model was stimulated by the recursive literature search model and can be solved using methods developed to deal with multiplicative stochastic processes. 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=02/12/14/013245;node=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.

In one incident Napoleon (incidentally, he was the military commander whose genius was questioned in War and Peace) 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
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.