**The h-index Paradox: Your coauthors have higher h-index than you**

Fabrício Benevenuto, Bruno Leite Alves, Alberto H. F. Laender

Department of Computer Science

Universidade Federal de Minas Gerais

Belo Horizonte, Brazil

H-index is a metric originally proposed to measure an individual’s scientific research output [2]. Its calculation is quite simple as it is based on the researcher’s set of most cited papers and the number of citations they have received. More specifically, a researcher has an h-index *h* if she published *h* papers that had received at least *h* citations. Thus, if a researcher has 10 papers with at least 10 citations, her h-index is 10.

Like any metric that attempts to summarize in a single number a complex and subjective evaluation, h-index has its limitations, including to be biased towards the authors scientific lifetime, not account for the number of coauthors in the publications and disregard the distinct citation patterns across different areas. Nevertheless, it became popular as it provides a notion of both quality and quantity of a researcher’s scientific output in a simple and easy to compute metric.

Therefore, researchers are tempted to evaluate themselves based on h-index. Systems like Google Scholar[[1]](#footnote-1) and ArnetMiner[[2]](#footnote-2) help researchers to track their publication impact and coauthors, as well as to maintain their profiles, where the h-index is stamped like a medal of honor.

Thus, it is reasonable to suppose that researchers may use their coauthors h-indexes as a way to infer whether they, themselves, have an adequate h-index in their areas or within a department or university. For instance, if you search on Google Scholar for a piece of your email (e.g., dcc.ufmg.br) you can obtain a ranking of your department colleagues according to their citation numbers.

There is nothing wrong with this. It is actually fun to browse others profiles to get a sense of how we are in comparison with other colleagues, especially our coauthors. However, this article seeks to show that this kind of comparison may lead to a classical paradox of sociology. We show that if an individual compares her h-index with the h-index of her coauthors, she might feel below the average.

We name this phenomena as the “The h-index Paradox” and present empirical results and sociological theories to support our arguments. We show that the mean h-index of a researcher’s coauthors is usually greater than her own h-index. We further explore the reasons behind this type of paradox. Next, first we briefly discuss how we estimate the h-index for researchers from distinct computer science scientific communities.

**Estimating the h-index of researchers and their coauthors**

In order to estimate the researchers’ h-indexes and compare them with the h-indexes of their coauthors, we need to be able to (1) identify the coauthors of a large set of researchers and (2) estimate the h-index of these researchers and their coauthors.

To construct the coauthorship network of computer science researchers of different areas, we gathered data from DBLP[[3]](#footnote-3), as it offers its entire database in XML format for download. We gathered this data for those researchers who published in the flagship conferences of 24 ACM SIGs (Special Interest Groups) [1]. Here, however, we focus on analyzing the researchers’ h-index in eight of them: SIGDOC, SIGCHI, SIGIR, KDD, SIGCOMM, SIGGRAPH, SIGMETRICS, POPL, and SIGMOD.

There are multiple tools that measure the h-index of researchers, out of which Google Scholar is today the most prominent one. However, to have a profile in this system, a researcher needs to sign up and explicitly create her research profile. In a preliminary collection of part of the profiles of the DBLP authors, we found that less than 30% of these authors had a profile at Google Citations [1]. Thus, this strategy would largely reduce our dataset.

To divert from this limitation, we used data from the SHINE (Simple HINdex Estimator) project[[4]](#footnote-4) to infer the researchers' h-index. SHINE provides a website that shows the h-index of almost two thousand Computer Science conferences. They crawled Google Scholar, searching for the title of papers published in these conferences, which allowed them to effectively estimate the h-index of the target conferences based on the citations computed by Google Scholar. Although SHINE only allows one to search for the h-index of conferences, the SHINE developers kindly allowed us to access their dataset to infer the h-index of researchers based on the conferences they crawled.

As SHINE does not track all existing Computer Science conferences, researchers' h-index might be underestimated when computed with this data. To investigate this issue, we compared the h-index of a set of researchers with a profile on Google Scholar with their estimated h-index based on the SHINE data. For this, we randomly selected 10 researchers for each of the ACM SIG’s flagship conferences and extracted their h-indexes from their Google Scholar profiles. In comparison with the h-index we estimated from SHINE, the Google Scholar values are, on average, 50% higher, but they are highly correlated (Pearson's correlation coefficient is 0.85), which indicates that researchers have proportional h-index estimations in both systems [1].

**Comparing the h-index of a researcher with her coauthors**

Having accurately estimated the h-index of each researcher, we can compare their h-index with her coauthors’.

Table 1 shows the fraction of authors with h-index smaller than the average of their coauthors for the eight conferences we consider. We can note that even focusing in authors that published in flagship conferences of ACM SIGs, the fraction of authors that might be below the average is quite high, varying from 69% to 81%. When we look at the percentage of authors with at least one coauthor with higher h-index than hers, the numbers are higher than 90% for most of the conferences.

**Figure 1:** Comparison results of one’s h-index with coauthors.

These results are quite related to a known phenomenon in social networks, called the Friendship Paradox [3, 4], which says that your friends in a social network have more friends than you on average. Although, coauthorship networks usually exhibit the same properties of many social networks [6, 7], including small-world properties and node degree distribution, this cannot explain alone the h-index Paradox as the Pearson correlation coefficient of an author’s h-index and her number of coauthors is 0.36, which is positive but not so high, suggesting that collaborating with a high number of researchers do not necessarily results in a higher h-index. Nevertheless, the important point here is not related to the reasons behind this paradox, but its consequences.

**Consequences of the Paradox**

The easy comparison among peers that today’s systems offer might accentuate the existing pressure that exists for publication as one might tend to feel pressured by the sensation of being below coauthors. Humans are natural competitors, not only in research, but in several aspects of life. The problem here is just an instantiation of a sensation that occurs in different scenarios and even culminated in an expression that is common to different languages and cultures: “The neighbor’s grass is always greener on the other side”.

Competition in science can be good as it gives extra motivation to researchers work hard, but it also may lead to undesirable scenarios, especially when researchers cross ethical boundaries, tempted to obscure experiments details or even fraud them to “sell” better their publications. A recent work on the field of Medicine has evaluated 788 retracted articles [8] and focused on 180 of them, which involves experiments with patients. From these, 70 were retracted by fraud and 110 by mistakes. These articles received more than 5,000 citations and involved thousands of patients. On the computer science perspective, retractions are not so common, but it is not hard to accept that there exist inaccurate reported results or even wrong code implementations published in major computer science venues.

Even without crossing any ethical line, it is natural that some individuals may try to adapt research strategies to bias their own scores towards a better h-index or citation count. The term “salami science” was used to designate when research results are split in pieces of publication to increase publication count [5]. But there is a price to pay when applying this kind of strategy. When a researcher slice results in minimal publishable pieces to produce more papers, she might even increase h-index as well, but she loses the chance to publish a seminal work, able to make people remember her for her important contribution. Potentially, a seminal work may have the power to indirectly increase her h-index.

There are many other strategies that one can think of, including self-citations in papers and forced citations to colleagues, which also comes with a price. For instance, suppose a researcher have 60% of its citation count as self-citations and Google scholar starts to explicitly show this percentage on the researcher’s profile. Then, her strategy would be clearly revealed.

What the h-index paradox says is that researchers might feel bellow the average in comparison with coauthors, thus, being tempted to play a sort of a game to increase scores. Playing this game may have a price and the rules of the game might change without any previous notice. The best strategy might be simply to not play any game. The goal is to advance the state of the art and produce high quality science. The output numbers should come naturally.

**References**

[1] Alves, B. L., Benevenuto, F., and Laender, A. H. F. The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the* *International Conference on World Wide Web (Companion Volume)*. ACM Press, New York, NY, 2013, 649-656.

[2] Hirsch, J. E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Science 102, 46 (2005), 16569–16572.

[3] Hodas, N. O., Kooti, F., and Lerman, K. Friendship Paradox Redux: Your Friends Are More Interesting Than You. *arXiv preprint arXiv:1304.3480* (2013).

[4] Feld, Scott L. "Why your friends have more friends than you do. *American Journal of Sociology 96*, 6 (1991), 1464-1477.

[5] Hoit, J. D. Salami science.  *American Journal of Speech-Language Pathology*  *16*, 2 (2007), 94.

[6] Huang, J., and Li, J. Collaboration Over Time: Characterizing and Modeling Network Evolution. In *Proceedings of the International Conference on Web Search and Web Data Mining*. ACM, New York, NY, 2008, 107-116.

[7] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. ACM Press, New York, NY, 2007, 29-42.

[8] Steen, R. G. "Retractions in the medical literature: how many patients are put at risk by flawed research?".  *Journal of Medical Ethics* *37*, 11 (2011), 688-692.

----------------------------------------------------------------------

**Viewpoints** articles should consist of up to 1,800 words, include a small number of references (generally, no more than 10), and be submitted to <http://cacm.acm.org/submissions>.

<http://cacm.acm.org/about-communications/author-center/author-guidelines#2.3.2>

1. http://scholar.google.com [↑](#footnote-ref-1)
2. http://arnetminer.org [↑](#footnote-ref-2)
3. <http://www.informatik.uni-trier.de/~ley/db/> [↑](#footnote-ref-3)
4. <http://shine.icomp.ufam.edu.br/> [↑](#footnote-ref-4)