ON HOUSE RENOVATION AND COAUTHORING: TRICKS OF THE TRADE TO BOOST YOUR H-INDEX

Roberto Piazza – Department CMIC 'Giulio Natta' – Politecnico di Milano – Italy – DOI: 10.1051/epn/2015103

Suppose that your house needs some restoration, and that you call a master mason asking for an estimate. If the mason replies at once that he will quote 1000 € for himself, plus 500 € for each helper apprentice, you will likely be puzzled, if not annoyed. Surely you have good reasons to complain, reasoning that the job you ask for should be remunerated with a fixed amount, irrespective of the number of labourers involved. Yet, this is not a criterion that we usually apply when evaluating the CV of an applicant for an academic position or for a grant.

We may examine the number of papers the applicant has written, where they have been published, or how many citations they have obtained. More recently, we would surely check the Hirsch h-index [2,3], or exploit more sophisticated indicators. Rarely do we look for the extent of coauthoring: a good paper is a good paper and, in terms of the applicant prestige, it is often regarded to be equally valuable regardless of whether it is signed by one, five, or two hundred coauthors. Possibly, if the applicant is the first author, who presumably did the hard job, or the last one, usually the lab "master mason", you may grant her or him an additional bonus. But that's all. After all, recovering quantitative information of this kind from search services like ISI or Scopus, even something simple as the average number of coauthors per paper, is not immediate (just try!).

Suppose, however, that the mason refutes your argument by claiming that the more people do the job, the better it comes out. You may be skeptical, but you will not easily come out with general abstract arguments in favour or against such a claim. Like a cosmologist who

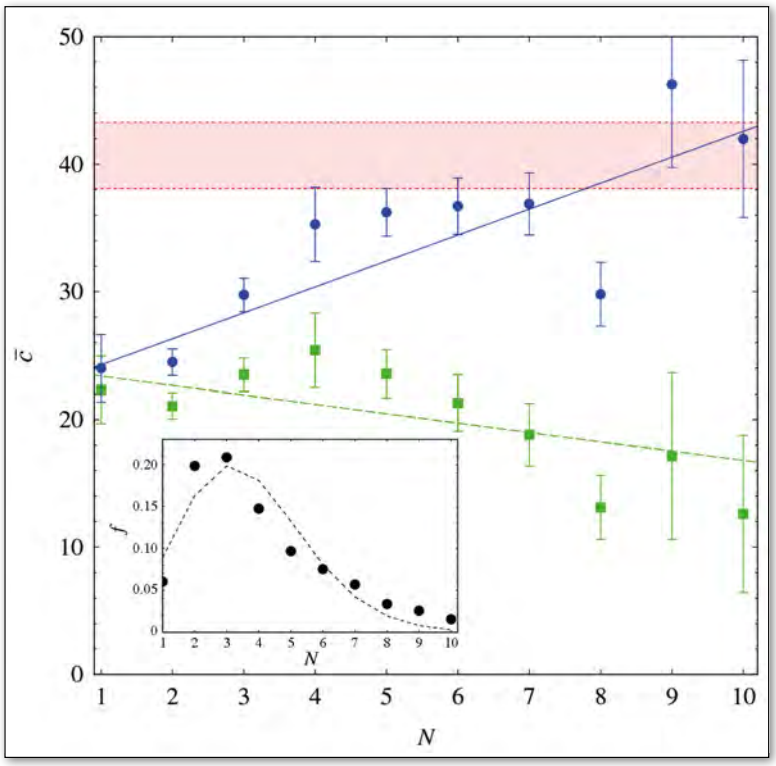
▲ Cartoon by W. Drenckhan, coauthored by D. Weaire and R. Piazza

has a single Universe to investigate, you have just this house to test, and relying on repetitive trials is out of question. In the scientific community, however, grounding discussions about coauthoring are not uncommon. Some colleagues argue that, yes, discouraging excessive coauthoring is probably sensible, but that a penalty consisting in simply dividing the citations of a given paper by the number N of authors is probably excessive. So they suggest using various sublinear forms of rescaling, such as dividing by $\sqrt{N}$, usually on the basis of some kind of *a priori* reasoning. Some others (mostly experimentalists), counter that being able to build up a collaboration network is a virtue that should be acknowledged, hence no scaling should be applied if N is, say, smaller than 5 or 10. When questioned, several high-energy experimental physicists even let the matter drop at once, branding talks of this kind as absurd. The fact is, in contrast to the former case, we *do* have a sensible, albeit not perfect way to quantify how much coauthoring impacts on the recognition of a publication by looking at the total number of citations it has received after some years. Faithful to the experimentalist's motto "In God we trust, all others must show data", let us then try and get some figures [1].

Coauthoring and citations.

Since I am addressing an audience of physicists, I shall focus on Physical Review Letters (PRL), still a reference journal for our community. I have considered the number of citations obtained in the first 6 years, according to ISI Web of Knowledge (WoK), by all manuscripts published in Physical Review Letters in 2007, which amounts to about 3700 records, including comments, but not replies and corrections. I sorted these papers in groups on the basis of the number of authors, and evaluated the average and standard deviation of the number of citations c for each group. A first striking evidence from the results, shown in Figure 1, is that c grows by a mere factor of two when N increases from 1 to 10, namely, just a little more than 8 % for each additional author. Equally surprising is that, as clearly evidenced by the purple band in Figure 1, very large collaborations do not seem to yield, on the average, a much greater impact on the scientific community. In other words, if we "reward" each author just on the basis on the total number of citations he/she has obtained, we are likely to make a big gift to those masons used to work in large groups.

Nevertheless, a moderate increase with N of the "acknowledged value" of a publication seems to be present. At least, if we neglect self-citations. Quantifying the latter for each single record is hard, and the WoK is surely not of great help. Just to get a rough figure, I then simply considered the average fraction of self-citations for those authors (about 150) of the 5% most cited papers who have got an ISI Author Identifier, which turns out to be 0.07 ± 0.01. If we then assume that each of the coauthors contributes to the total number of a citations of a given paper with 7% of self-citations, we may subtract this "spurious" contribution by substituting $c \rightarrow (1\text{-}0.07N)c$. This is of course questionable, since several papers have been probably cited by more than one coauthor, hence the contribution of self-citations is likely to be overestimated. Yet, the result is rather impressive, for the net data obtained this way (squares in Figure 1) even show a slight apparent *decrease* with N. We may conclude that the scientific impact of a paper is roughly independent of N.

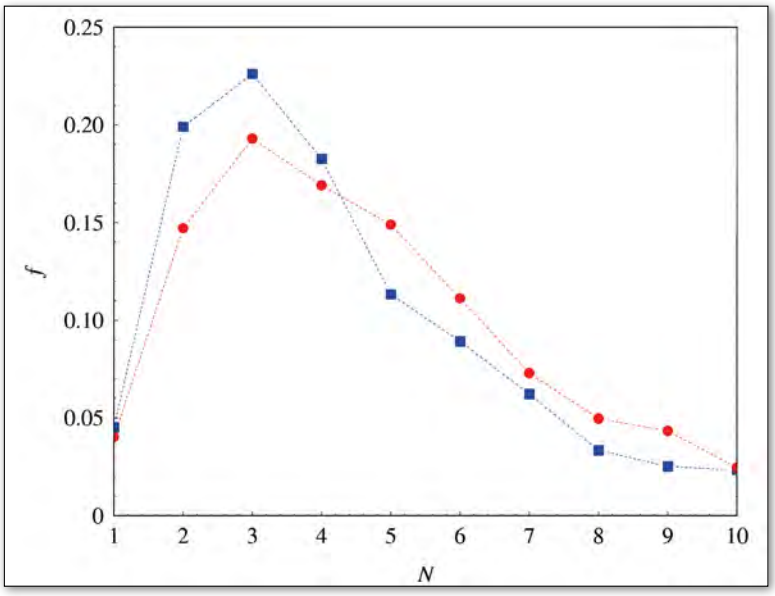


▼ **FIG 1:** Average number of citations $\bar{c}$ versus the number N of authors at the end of 2012 for the manuscripts published in PRL in 2007 with $N \leq 10$ (blue dots). The full line is a linear fit with slope (0.08 ± 0.02) N. The purple band shows the number of citations (within $\pm 1\sigma$) of the papers with $N > 10$, which are about 8 % of the total. When self-citations are tentatively removed by rescaling $\bar{c}$ by a factor $(1+0.07N)^{-1}$, the corrected data point (green squares) show no significant change, or even a slight decrease, with N. Data are obtained from a set of about 3400 records, with the distribution shown in the inset, and compared with a Poisson distribution having the same mean (dotted line).

Coauthoring and excellence.

Yet, the fact that a multi-authored article is *on the average* not cited more than a paper originating from a single small group may not be the real issue. Perhaps, young scientists (but also experienced group leaders) long for collaborations because they believe this gives more chances to a publication of entering the restricted heaven of *excellent*, outstanding papers. Besides, the story may not be the same for different research areas in physics, an aspect which is not captured by a crude analysis of the total number of citations. Once again, let's have a look at real data. Things are relatively easy if one considers only those papers that are awarded a little "gold cup" in the WoK because, according to the Essential Science Indicators (ESI), they score within the 1% more cited papers, in a given year, within a subfield of a discipline. Hence, I have considered *all* papers published by PRL within the past decade (2004-2013) that should be regarded as "excellent" according to the ESI. Let us first discuss the results obtained for those

papers with $N \leq 10$, which amount to about 1900 articles over a total of more than 38000 papers published by PRL in the decade. Their relative frequency distribution as a function of the number of authors, shown in Figure 2 with red dots, is of course far from being uniform, just because such is the total number of submitted and published papers (see the inset in Figure 1). To compare consistently with the latter, I have considered an equal number of papers, selected by merely sorting them, for each year, in terms of publication date. Which means, basically, at random. The results I obtained are shown by the blue squares in Figure 2. Although statistically significant differences *can* be spotted, so that increasing the number of authors seems to give a slightly larger chances of making a very successful paper, these are minimal: for instance, the average number of authors in the distribution of excellent papers (about 4.5) is very close to the value obtained for the distribution of randomly selected PRL papers (about 4.1). PRL articles with $N > 10$ account for about 16% of the ESI selected papers, whereas they amount to only 10% of the total number of published papers. Sticking to our analogy, if the master mason summons a large group of apprentices - say, 100 - the chance they make a superior job increases by 60-70% with respect to the case of the master working alone: remember, however, that we pay *fifty* times more. Summing up, I am prone to conclude that the "merit" of a scientific publication, as judged by the number of citations it obtains, or by its chance of "scoring at the top", does not substantially depend on N. Hence, in the absence of further information on the role played by each author (of the kind provided for instance by several biological or medical journals), credit should be shared in equal parts by all coauthors.



▲ FIG 2: Frequency distribution versus the number N of coauthors of the papers published in PRL in the decade 2004-2013 considered as "outstanding" according to the Essential Science Indicators (red dots), compared with the distribution of an equal number of published papers, randomly selected according to the publication date (blue squares).

Does quality require quantity?

In bibliometric assessments, taking into account the former "profit sharing" considerations in detail is not trivial [4]. A crude but reasonable approach is simply rescaling the total number of the citations of a scientist by the *average* number of authors of her/his papers, or, in the case of the h-index, by the average number of authors of her/his h most cited papers [5,6]. (A more sophisticated approach has been taken by Hirsch himself [7], who introduced a so-called "$\hbar$-index", roughly defined as follows: one of your papers contributes to your $\hbar$-index if it contributes also to the $\hbar$-index of *all* your co-authors.)

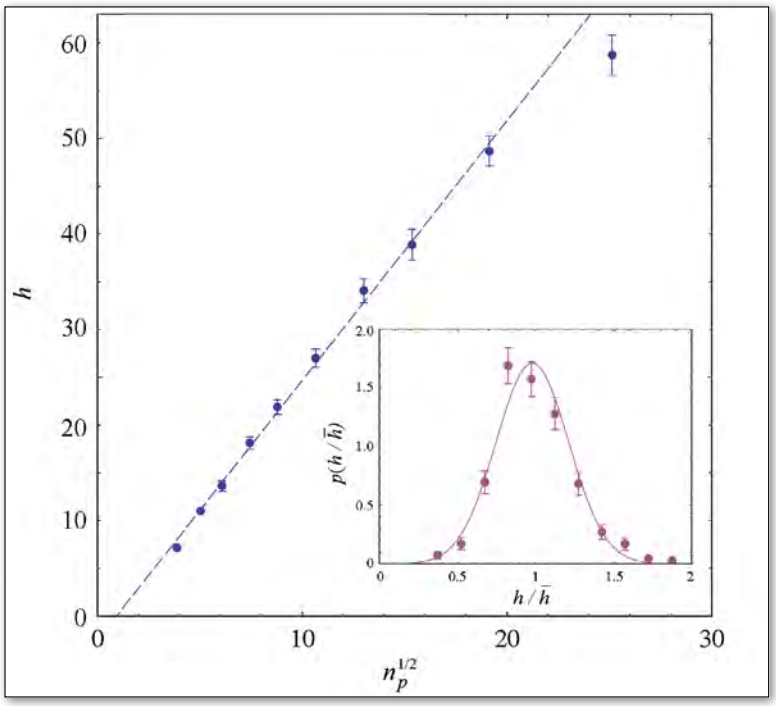
A further excursus on the h-index is, however, appropriate. Because it is so easy to evaluate, but more so because of its statistical robustness, the Hirsch index has rapidly ascended the throne of bibliometrics as a single number summarizing the success of a scientist: I myself must confess of having been a fan, almost a zealot of this brilliant, straightforward approach since it was originally proposed. Yet, how much additional information does the h-index *really* convey? We may reasonably expect h to scale with $\sqrt{c}$ (at least, $\sqrt{c}$ is obviously an upper bound for h). But is there any relation between h and the total number of papers an author has published? To this aim, I have considered the 10% most cited papers published in PRL in 2012, examining (manually) the individual citation reports of all those authors, 470 in total, who appear to have an ISI Author Identifier. As can be reasonably expected, if n_p is the total number of papers an author publishes, the ratio h/n_p (which we may regard as a kind of "success ratio") rapidly decreases with n_p. Actually, Figure 3 shows that h is quite well fitted by a linear dependence on $\sqrt{n_p}$, except for $n_p \geqslant 400$, where some saturation may occur. What is really surprising is the very limited dispersion of the data around the mean. As a matter of fact, the ratio of the actual h-index for each individual author to the value $\bar{h}(n_p)$ obtained from the fit has an approximately Gaussian distribution, with a standard deviation $\sigma = 0.23$.

In simple words, this means the following: tell me the total number of papers you have published, and I'll predict your h-index within 20-30 % accuracy. More seriously, this result cast doubts on the amount of novel information the h-index carries *per se*, besides a simple reshuffling of basic information about the total scientific productivity of an author. Notice that even more refined bibliometric parameters like the "contemporary h-index" [8], which suitably takes into account the total duration and trend of the scientific production of an author, would not perform much better. (In fact, provided that these

"The merit of a scientific publication, as judged by the number of citations it obtains, does not substantially depend on the number of authors"

general observations are confirmed by testing a much larger and varied sample besides the selected one I have considered, a more meaningful bibliometric parameter would actually be the fractional *deviation* $\delta h = h/\bar{h}(n_p) - 1$). The observed correlation between h and n_p together with the basic independence of the value of a scientific paper on N, could be particularly deleterious for the community of experimental high-energy or nuclear physicists, whose h-indices, besides being typically larger than the average, have a distribution with a quite smaller relative standard deviation $\sigma h/\bar{h}$ [1]. Hence these authors form a rather homogeneous group in term of their overall "scientific success", which of course makes it harder to discriminate among different scientists on the bases of the h-index.

To conclude, I am surely not claiming the little evidence I dug out to be conclusive or comprehensive: this little *divertissement* should not be taken too seriously, for any sound conclusions must be corroborated by a much more extensive and rigorous statistical analysis, which could easily be performed by appropriate organizations such as ISI or Scopus. My aim has simply been to try and shift the discussion about the impact of coauthoring from abstract reasoning to real data analysis. Nevertheless, the former observations surely lead me to two considerations. First, in the future I would not like to take part in committees where hiring or funding of young scientists is made only on bibliometric bases, renouncing to the pleasure of interviewing, even shortly, the candidates. Second, I came to believe that no bibliometric approach to hiring and promoting, however refined, will ever ensure a real improvement of our academic institutions, unless there are ultimate motivations *to long* for scientific quality. And, at least within some national communities, this should not be necessarily taken for granted. ■



▼ **FIG 3:** Hirsch index h as a function of the square root of the number of published papers n_p, for a set of 470 scientists co-authoring the 10% top cited papers published by PRL in 2012. The quantity $\bar{h}$ is derived from n_p according to the fit $\bar{h}(n_p) = (2.72 \pm 0.05)\, n_p^{1/2} - (2.5 \pm 0.5)$. The inset shows the frequency distribution of the quantity $h/\bar{h}(n_p)$ for the whole set of investigated authors and fitted with a Gaussian with standard deviation $\sigma = 0.23$.

Acknowledgements

The original version of this paper [1] has been prepared while I was in Cambridge as a visiting professor: I thank Pietro Cicuta for having given me the chance to work in such a stimulating environment. I also took pleasure from discussing these issues with several colleagues, and in particular with Wilson Poon, a scientist well on the right (in both senses) side of the Gaussian in Figure 3. I am finally extremely grateful to Wiebke Drenckhan, who, besides bringing Ref. [1] to the attention of EPN, contributed to this paper with the splendid opening cartoon.

About the author

Roberto Piazza is full professor of Condensed Matter Physics in Politecnico of Milan, after having got his PhD from the University of Pavia and having worked as a research associate at the University of Pittsburgh. He has also being invited professor at the École Normale Supérieure in Lyon and Leverhulme visiting professor at the Cavendish Lab in Cambridge.

His research has been mostly devoted to the experimental investigation of soft matter systems ranging from colloidal suspensions to surfactant and protein solutions. Currently, he is Section Editor for liquids, soft matter, and biological physics of the Journal of Physics Condensed Matter.

References

[1] A preliminary version of this paper, containing some additional technical details, can be freely download from the arXiv:1307.5647 [physics.soc-ph].

[2] J. E. Hirsch, An index to quantify an individual's scientific research output, *Proc. Natl. Acad. Sci. U.S.A.* **102**,16569 (2005).

[3] P. Ball, Index aims for fair ranking of scientists, *Nature* **436**, 900 (2005).

[4] As a matter of fact, rescaling by the number of authors of each single publication is automatically feasible using software like "Publish or Perish". Provided of course you fully trust the Google Scholar database on which this programme is based: just a matter of taste.

[5] P, D. Batista, M. G. Campiteli, O. Kinouchi, Is it possible to compare researchers with different scientific interests?, *Scientometrics* **68**, 179 (2006)

[6] Several drawbacks of this simple approach can be eliminated by introducing a "fractional count" of the papers, see M. Schreiber, A modification of the h-index: The hm-index accounts for multi-authored manuscripts, *Journal of Informetrics* **2**, 211 (2008)

[7] J. E. Hirsch, An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship, *Scientometrics* **85**, 741 (2010)

[8] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, Generalized Hirsch h-index for disclosing latent facts in citation networks, *Scientometrics* **72**, 253 (2007)