Scientometrics (2011) 89:291–299
DOI 10.1007/s11192-011-0451-5

Credit where credit's due: accounting for co-authorship in citation counts

Richard S. J. Tol

Received: 9 May 2011 / Published online: 16 July 2011

Abstract I propose a new method (Pareto weights) to objectively attribute citations to co-authors. Previous methods either profess ignorance about the seniority of co-authors (egalitarian weights) or are based in an ad hoc way on the order of authors (rank weights). Pareto weights are based on the respective citation records of the co-authors. Pareto weights are proportional to the probability of observing the number of citations obtained. Assuming a Pareto distribution, such weights can be computed with a simple, closed-form equation but require a few iterations and data on a scholar, her co-authors, and her co-authors' co-authors. The use of Pareto weights is illustrated with a group of prominent economists. In this case, Pareto weights are very different from rank weights. Pareto weights are more similar to egalitarian weights but can deviate up to a quarter in either direction (for reasons that are intuitive).

Keywords Citations · Co-authors · Pareto distribution

Introduction

Papers with multiple authors pose a problem to scientometric analysis. Who deserves the credit? There are three common solutions to this (Abbas 2011). The first and perhaps most common approach is to ignore the problem and let all co-authors take full credit. Bad incentives are the result: Authors may add each other to their papers even without a

R. S. J. Tol (✉)
Economic and Social Research Institute, Dublin, Ireland
e-mail: richard.tol@esri.ie

R. S. J. Tol
Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands

R. S. J. Tol
Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands

R. S. J. Tol
Department of Economics, Trinity College, Dublin, Ireland

contribution.[1] The second solution—"egalitarian weights"—is to share the credit equally between the co-authors (Batista et al. 2006; Ellison 2010; Schreiber 2008). Essentially, the analyst claims to have no information about who contributed most. The third solution—"rank weights"—is to share the credit based on the order of the authors (Hagen 2009; Hodge and Greenberg 1981; Sekercioglu 2008; Zhang 2009). This is entirely ad hoc, and conventions on the order of authors differ greatly between disciplines. None of these rules are satisfactory. In this paper, I present a method to apportion credit in an objective manner using readily available data.[2]

The idea is straightforward. Suppose that a mediocre researcher and a star jointly write a paper. The paper is widely cited. Surely, most of the credit should go to the star.

While simple, the idea cannot be implemented without defining the relative stardom of the two authors. Again, I adopt a simple approach. Based on the citation record of the two researchers, I find the probability that a paper by them is cited N times. By definition, the star would have a higher probability of N citations (if N is large) than the mediocre researcher. The relative credit is proportional to the relative probabilities.

In the next section, I formalize this, defining Pareto weights. "The data" section presents the data used to illustrate the proposal. "Results" section discusses the results. "Discussion and conclusion" section concludes.

A method for attributing citations

Consider S scholars who published papers that were cited $C_{i,s} > 0$ times. For convenience, we disregard uncited papers. The Pareto distribution (Pareto 1896) is often used to describe the number of citations (Egghe 1987, 1991, 1998, 2005):

$$f\left(C_{i,s}\right) = \frac{\mu\alpha_s}{C_{i,s}^{1+\alpha_s}} \tag{1}$$

We set $\mu = 1$ so that we allow for any number of citations.[3] The maximum likelihood estimate for the Pareto index α is

$$\alpha_s = \frac{n_s}{\sum_{i=1}^{n_s} C_{i,s}} \tag{2}$$

Now consider a paper l that is cited C times and that is co-authored by scholars s and t. Each scholar is allocated a share w of the citations according to

$$w_{l,s} = \frac{\alpha_s C^{-\alpha_s-1}}{\alpha_s C^{-\alpha_s-1} + \alpha_t C^{-\alpha_t-1}}; w_{l,t} = \frac{\alpha_t C^{-\alpha_t-1}}{\alpha_s C^{-\alpha_s-1} + \alpha_t C^{-\alpha_t-1}} \tag{3}$$

Obviously, $w_s + w_t = 1$; $w_s = w_t = 1/2$ if and only if $\alpha_s = \alpha_t$. Equation 3 has that scholar s receives the greater credit for the joint publication if the number of actual citations is more in line with her citation record. Equation 3 readily generalizes to multiple authors. I refer to w_s as the Pareto weight of author s.

[1] Note that collaborative research tends to be cited more often (Levitt and Thelwall 2010).

[2] As an alternative, one could rely on survey data (Vinkler 1993).

[3] Strictly, the Pareto distribution is defined on real numbers $C > \mu$. This is convenient if citations are shared between co-authors (as done below). For now, one can think of $f(C)$ as $F(C + 0.5) - F(C - 0.5)$.

There are two problems. First, the joint publication is part of the citation record that is used to estimate the Pareto index α. Therefore, the joint publication is used to assess itself. This would be avoided if the joint publication is excluded from Eq. 2. For scholars with a large number of cited papers, this does not make much of a difference. Nitpickers are free to use $\alpha_{s,\{1\}}$.

The second problem is more substantial. Equation 2 uses the full number of citations. Equation 3 allocates only a fraction of the citations to scholar s. That is, Pareto weights change the citation record. The method is internally inconsistent. In order to solve this, redefine Eqs. 2–3 as the 0th iteration. In the mth iteration,

$$\alpha_s^{(m)} = \frac{n_s}{\sum_{i=1}^{n_s} w_s^{(m-1)} C_{i,s}} \tag{4}$$

and

$$w_s^{(m)} = \frac{\alpha_s^{(m)} C^{-\alpha_s^{(m)}-1}}{\sum_t \alpha_t^{(m)} C^{-\alpha_t^{(m)}-1}} \tag{5}$$

The number of iterations should be such that $w^{(m)} \approx w^{(m-1)}$.

There is a practical problem with the above proposal. A scholar's corrected citation record depends on the citation record of everyone she has ever published with, and on everyone they have ever published with, and so on. Equations 4–5 can therefore only be approximated.

The data

I illustrate the above proposal with the case of Andrei Shleifer, a professor of economics at Harvard University. Although only 50 years old, Shleifer tops the IDEAS/RePEc life-time achievement ranking of all economists.[4] Shleifer won the John Bates Clark Medal and is likely to win the Nobel Prize. He has a limited number of long-term collaborators, which eases data collection.

I collected the publication and citation record of Andrei Shleifer, his 36 collaborators, and 4 of the collaborators of his closest collaborators. I did this at Easter 2011, using Scopus[5] as the source of data.

Table 1 lists the names, numbers of (cited) publications, numbers of citations, and the Hirsch (Hirsch 2005) and Pareto indices (Eq. 2). Table 1 also gives the Shleifer-number: 0 for Shleifer, 1 for his coauthors, 2 for his coauthors' coauthors.[6] Table 1 contains a relatively small (41) but very diverse group of scholars. There are scholars who are generally considered to be world class, former post-docs who left academia, and everything in between. This is appropriate for illustrating the proposal of "A method for attributing citations" section.

[4] http://ideas.repec.org/top/top.person.all.html.

[5] http://www.scopus.com/home.url.

[6] The current author's Shleifer number is 3.

Table 1 Selected characteristics of the authors in the sample: number of papers, number of cited papers, number of citations, average number of citations (per cited paper), Hirsch index, Pareto index, and Shleifer number

Authors	Papers	Cited	Citations	Ave Cit	Hirsch	Pareto	Shleifer
Shleifer	81	77	16385	212.8	50	0.2346	0
Glaeser	89	86	5175	60.2	33	0.3206	1
Aghion	87	75	2627	35.0	26	0.3624	1
Lopez-de-Silanes	30	28	9277	331.3	24	0.2037	1
Markusen	54	49	2399	49.0	24	0.3381	2
Blanchard	55	41	2228	54.3	24	0.2901	1
la Porta	26	25	9325	373.0	23	0.1920	1
Lakonishok	28	28	2246	80.2	22	0.2646	1
Vishny	27	25	10758	430.3	21	0.1970	1
Johnson	35	33	3568	108.1	20	0.2885	1
Djankov	52	47	2918	62.1	19	0.3478	1
Morck	32	27	2287	84.7	15	0.3336	1
Scheinkman	34	27	1388	51.4	15	0.3428	1
Rutherford	56	51	715	14.0	15	0.4865	2
Murphy	26	25	2053	82.1	14	0.3320	1
Mullainathan	31	25	1405	56.2	14	0.3315	1
Treisman	32	29	825	28.4	14	0.4150	1
Hart	27	24	1285	53.5	13	0.3539	1
Wurgler	17	14	1065	76.1	12	0.2684	1
Barberis	12	12	1429	119.1	11	0.2448	1
Kaufmann	16	16	697	43.6	11	0.3259	1
Mulligan	26	22	541	24.6	11	0.3969	1
Frye	15	13	327	25.2	8	0.4055	1
Cahuc	33	21	234	11.1	8	0.5585	1
Wolfenzon	8	7	459	65.6	7	0.2584	1
Burkhart	9	8	435	54.4	7	0.3091	1
Panunzi	9	8	425	53.1	6	0.3185	1
Boyko	5	5	484	96.8	5	0.2606	1
McLiesh	4	4	201	50.3	4	0.2814	1
Algan	16	9	46	5.1	4	0.8008	1
Gennaioli	9	5	37	7.4	4	0.5521	1
Nenova	3	3	221	73.7	3	0.2745	1
Pop-Eleches	7	4	81	20.3	3	0.5018	1
Ponzetto	4	4	38	9.5	3	0.5257	1
Volokh	6	4	30	7.5	3	0.7174	1
Botero	2	2	212	106.0	2	0.3105	1
Hay	2	2	74	37.0	2	0.2772	1
Tsukanova	1	1	125	125.0	1	0.2071	1
Zamarripa	1	1	61	61.0	1	0.2433	2
Schwartzstein	1	1	10	10.0	1	0.4343	1
Moore	1	1	10	10.0	1	0.4343	2

Results

Figure 1 shows the Pareto index as a function of the Hirsch index (bottom panel) and as a function of the average number of citations per publication (top panel). The Pareto index is the inverse of the average of the natural logarithm of the citation number (see Eq. 2), but Fig. 1 shows that the inverse of the log of the average citation number is reasonable approximation. Figure 1 also shows that there is a relationship between the Pareto and Hirsch indices—a high number of highly-cited papers imply both a low Pareto index and a high Hirsch index—but that they measure different things—the Hirsch index disregards excess citations while the Pareto index does not.

Let us now turn the attention to the attribution of citations to joint papers to individual authors, focusing on Shleifer, the central author in the sample. The top panel of Fig. 2 shows the histograms of Shleifer's Pareto weights for his papers with one, two, three and four other authors. Shleifer did not publish in teams of six or more. The Pareto weight for single authored papers is, by definition, one. Figure 2 shows that the Pareto weights spread around the egalitarian weights ($1/n$ where n is the number of authors). Egalitarian weights are thus a reasonable approximation of Pareto weights. By implication, rank weights

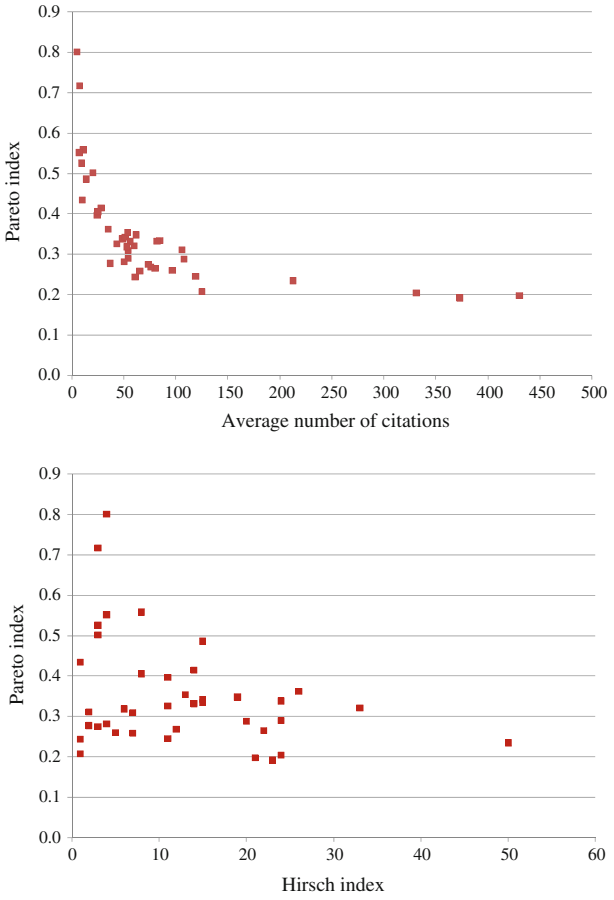

Fig. 1 The Pareto index versus the average number of citations per paper (*top panel*) and the Hirsch index (*bottom panel*) for the 41 scholars in Table 1

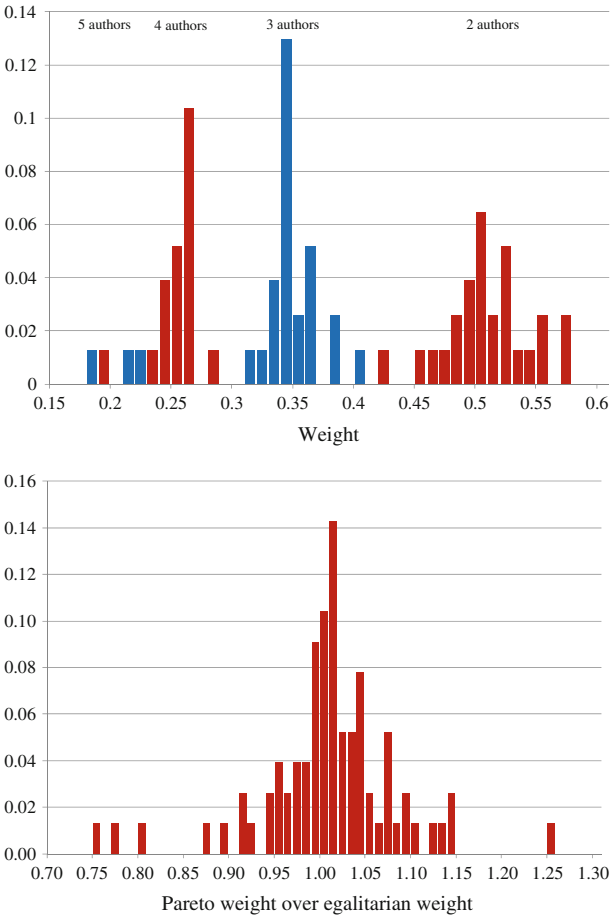

Fig. 2 The histograms of the Pareto weights for different numbers of authors (*top panel*) and the histogram of the ratio of the Pareto weights to the egalitarian weights (*bottom panel*)

(proportional to 1 for the first author, 1/2 for the second, …) are not. This is no surprise given the convention in economics to list authors alphabetically.

Shleifer receives a more-than-egalitarian weight for some papers, but one may be surprised that he receives less-than-egalitarian weight for other papers.[7] After all, Table 1 shows that he is more senior than any of his coauthors. The bottom panel of Fig. 2 confirms its top panel. The bottom panel shows the histogram of the ratio of the Pareto weights to the egalitarian weights. The histogram is centred around one (so that the egalitarian weights are a reasonable approximation). The distribution ranges from 0.75 to 1.25, that is, the egalitarian weight may be a quarter too high or too low (if one accepts the Pareto weight as the true weight).

Let us consider the two extreme cases of the bottom panel of Fig. 2 in order to develop some intuition about the Pareto weights. The ratio of Pareto to egalitarian weights is highest for a paper co-authored by Barberis et al. (1998). It was cited 503 times. This is extraordinary for Barberis (whose papers are cited 119 times on average), not so special for

[7] Note the difference with the $\hbar$ index (Hirsch 2010), which always gives full credit to the most senior author and gives either full or no credit to junior authors.

Table 2 Selected characteristics of the four core authors in the sample: number of cited papers (P); average number of authors (A); average number of citations ($C^{(0)}$) and after egalitarian ($C^{(E)}$) correction and Pareto correction ($C^{(1)}$, $C^{(2)}$); and Pareto index for all citations ($P^{(0)}$) and after egalitarian ($P^{(E)}$) correction and Pareto correction ($P^{(1)}$, $P^{(2)}$)

	P	A	Average citations per paper				Pareto index			
			$C^{(0)}$	$C^{(E)}$	$C^{(1)}$	$C^{(2)}$	$P^{(0)}$	$P^{(E)}$	$P^{(1)}$	$P^{(2)}$
Shleifer	77	2.8	212.8	72.5	71.8	74.0	0.2346	0.3027	0.3030	0.3022
Lopez-de-Silanes	28	3.6	331.3	89.9	91.5	89.8	0.2037	0.2728	0.2729	0.2734
La Porta	25	3.7	373.0	104.0	107.5	104.5	0.1920	0.2541	0.2533	0.2540
Vishny	25	3.0	430.3	139.9	146.4	144.5	0.1970	0.2501	0.2493	0.2496

Sheifer (whose papers are cited 213 times on average) and run-of-the-mill for Vishny (whose papers are cited 430 times on average). The egalitarian weights are one-third for each author. The Pareto weights are 15% for Barberis, 41% for Shleifer and 44% for Vishny.

At the other extreme lies a paper by Aghion et al. (2010). It was cited only four times.[8] This is exceptional for Shleifer (213 citations on average), not uncommon for Aghion (35 citations on average), common for Cahuc (11 citations on average) and as expected for Algan (5 citations on average). Therefore, the Pareto weights are 0.29 (Algan), 0.28 (Cahuc), 0.24 (Aghion) and 0.18 (Shleifer); the egalitarian weight is 0.25. This highlights another property of Pareto weights: Because a probability density function integrates to one, scholars with a high probability of a large number of citations have a low probability of a small number of citations. Pareto weights thus attribute a large share of the citations to highly-cited papers to highly-cited co-authors, and a small share of the citations to little-cited papers to highly-cited co-authors.

The above results are for the 0th iteration of the Pareto weights. See Eqs. 2–3. I computed the 1st iteration for Shleifer and his three core collaborators: La Porta, Lopez-de-Silanes and Vishny. There are seven papers with these four people as co-authors, and four of these papers are cited more than 500 times. There are another four papers by La Porta, Lopez-de-Silanes and Shleifer; nine papers by La Porta, Lopez-de-Silanes, Shleifer and others; seven papers by Shleifer and Vishny; and eleven papers by Shleifer, Vishny and others. All of Vishny's papers are co-authored by Shleifer; 80% of La Porta's papers; and 71% of Lopez-de-Silanes' papers. 49% of Shleifer's papers are with some or all of these core collaborators.

Table 2 repeats some of the characteristics from Table 1 and adds new ones for these four scholars. The Pareto weights allocate 34% of citations to Shleifer and Vishny, compared to 28–29% to La Porta and Lopez-de-Silanes. The latter two have lower Pareto indices and thus a greater probability of publishing highly-cited papers. However, Shleifer and Vishny tend to publish with fewer co-authors, and this effect dominates the difference in Pareto indices.

This effect is reinforced in the first iteration, in which the Pareto index is calculated for the attributed citations. The average number of citations and the Pareto index fall for each of the four authors (as they receive 100% or less of the citations). However, the Pareto index rises more for La Porta and Lopez-de-Silanes than for Shleifer and Vishny.

[8] Note that this is a recent paper. This issue is further discussed below.

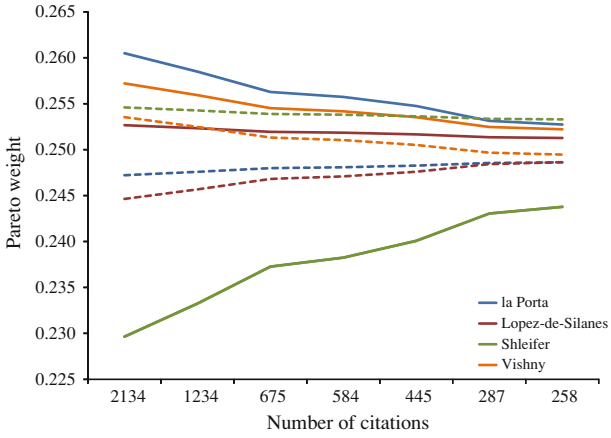

Fig. 3 The Pareto weights assigned to the four scholars in the 0th (*dashed lines*) and 1st (*solid lines*) iterations as a function of the number of citations that the papers received

Table 2 also shows the attributed citations and Pareto indices for the second iteration. Although the attribution again shifts in favour of Shleifer and Vishny, the differences with the first iteration are minimal. At least for this group of authors, the first iteration appears to be a reasonable approximation. Table 2 further shows that the egalitarian attribution is, at least in this case, a reasonable approximation (but not in all cases as shown in Fig. 2).

Figure 3 highlights the difference between the 0th and 1st iterations for the seven papers co-authored by the four core scholars. Three things stand out. Firstly, there is a change in the order of attribution. Whereas in the 0th iteration, the credit went to La Porta first, Vishny second, Lopez-de-Silanes third and Shleifer fourth; in the 1st iteration, Shleifer is first, followed by Vishny, La Porta and Lopez-de-Silanes. In both iterations, there is a difference with the egalitarian attribution (0.25)—but, noting the vertical scale of Fig. 3 (2.25–2.65), the difference is small.[9] Secondly, attribution varies less with the number of citations. This is because differences in the Pareto index matter more for higher citation numbers, and citations numbers are lower when shared between co-authors.

The above results are based on the number of citations per paper. This is a proper measure for the eventual impact of a scholar, but Shleifer is an active researcher and some of his papers were published too recently to amass a large number of citations (see above). Therefore, I repeated the analysis with citations per year—specifically, citations divided by 2012 minus the year of publication. Using this metric, 33.81% of citations per year are attributed to Shleifer. This compares to 33.75% of citations. In this case, therefore, citations and citation-rates yield indistinguishable results.

Discussion and conclusion

I propose an objective method to attribute citations to co-authors. The Pareto weight is based on the probability of observing a number of citations given the author's citation record. Assuming that citation numbers follow a Pareto distribution, there is a closed-form solution to compute the Pareto weight. However, one needs a few iterations and data on the

[9] There is a large difference with the standard rank attribution (Hagen 2009; Hodge and Greenberg 1981; Sekercioglu 2008). In that case, La Porta would be attributed 48% of the citations, Lopez-de-Silanes 24%, Shleifer 16%, and Vishny 12%.

scholar in question as well as on her co-authors and their co-authors. In the examples used in this paper, the Pareto weights attribute up to 25% more or less citations to an author than do equal weights. The Pareto weights are very different from rank-based weights.

In future research, it would be good to test the current proposal with other data. A longitudinal study would be particularly interesting. Over time, a scholar's publication and citation record changes. Using Pareto weights, the attribution of citations changes too. One could, of course, also consider alternative distributional assumptions, particularly when modeling citation-rates rather than citations (Fok and Franses 2007; Franses 2003).

References

Abbas, A. (2011) Weighted indices for evaluating the quality of research with multiple authorship. *Scientometrics*, *88*(1), 107–131.

Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and distrust. *Quarterly Journal of Economics, 125*(3), 1015–1049.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, 49*(3), 307–343.

Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martinez, A. S. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Egghe, L. (1987). An exact calculation of Price's law for the law of Lotka. *Scientometrics, 11*(1–2), 81–97.

Egghe, L. (1991). The exact place of Zipf's and Pareto's law amongst the classical informetric laws. *Scientometrics, 20*(1), 93–106.

Egghe, L. (1998). Mathematical theories of citation. *Scientometrics, 43*(1), 57–62.

Egghe, L. (2005). A characterization of the law of Lotka in terms of sampling. *Scientometrics, 62*(3), 321–328.

Ellison, G. (2010), *How does the market use citation data? The Hirsch index in economics*, Working Paper 3188, CESifo, Munich.

Fok, D., & Franses, P. H. (2007). Modeling the diffusion of scientific publications. *Journal of Econometrics, 139*, 376–390.

Franses, P. H. (2003). The diffusion of scientific publications: The case of *Econometrica*, 1987. *Scientometrics, 56*(1), 29–42.

Hagen, N. T. (2009). Credit for coauthors. *Science, 323*(5914), 583.

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

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics, 85*(3), 741–754.

Hodge, S. E., & Greenberg, D. A. (1981). Publication credit. *Science, 213*, 950.

Levitt, J. M., & Thelwall, M. (2010). Does the higher citation of collaborative research differ from region to region? A case study of Economics. *Scientometrics, 85*(1), 171–183.

Pareto, V. (1896). *Cours d'Economie Politique*. Lausanne: F. Rouge.

Schreiber, M. (2008). A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *Journal of Informetrics, 2*(3), 211–216.

Sekercioglu, C. H. (2008). Quantifying coauthor contributions. *Science, 322*(5900), 371.

Vinkler, P. (1993). Research contribution, authorship and team cooperativeness. *Scientometrics, 26*(1), 213–230.

Zhang, C. T. (2009). A proposal for calculating weighted citations based on author rank. *EMBO Reports, 10*(5), 416–417.