Patterns of Co-authorship: Statistical, Philosophical, and Sociological Implications

J Donald deB. Beaver

Williams College, Dept. of History of Science,Bronfman Science Center 117, 18 Hoxsey St.
Williamstown, MA 01267 - USA

Abstract
The vanishing single authored-paper seems to have hit a limit below which it will strongly resist further encroachment by collaboration or teamwork. That limit is about 1 paper in 10, but varies by field of science, being lower (~1 in 16) in chemistry, but higher in biology (~ 1 in 7). The prevalence of single authorship also varies between experimental (lesser) and theoretical (greater) research. The overall pattern of the growth of collaborative research now displays the shape of a logistic curve. Teamwork (6 or more authors) continues to grow in all fields at the expense of collaboration (5 or fewer authors), but the two patterns remain statistically distinct. The biological sciences came last to display the co-authorship pattern of teamwork.

The Single Author Redux: Has Collaboration Reached Saturation?
In 1967 and again in 1987, counts of singly and multiply-authored papers in Biological Abstracts, Chemical Abstracts, and Physics Abstracts (Science Abstracts), showed that the percentage of scientific papers that were co-authored was following an upward trend that looked like the end of the century would signal the demise of research formally attributable to a single investigator. Now some new counts from those same abstracting journals indicate that the individual researcher continues to publish about 10% of all papers.

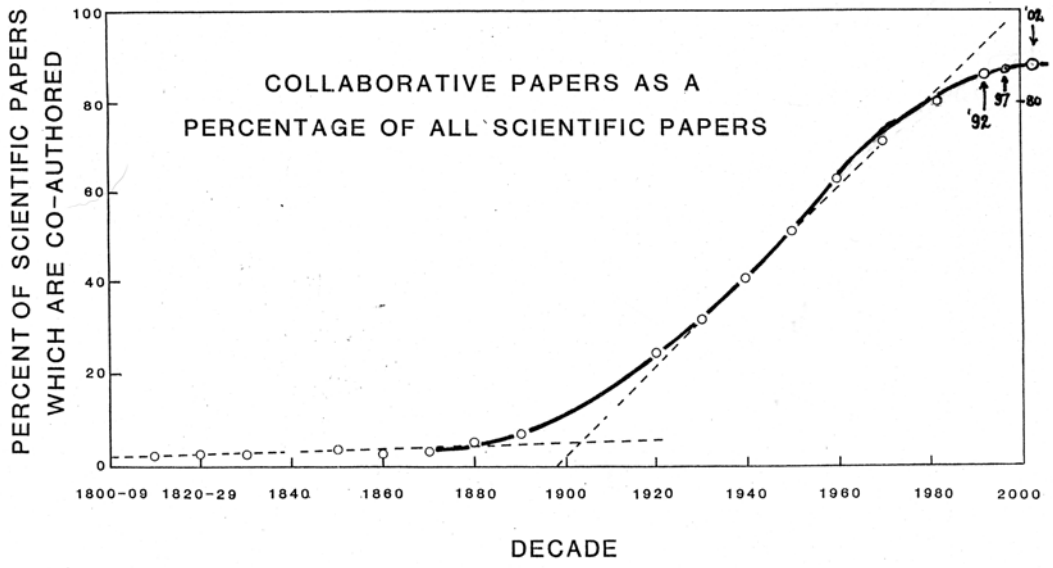

Figure 1.

This result is interesting for a number of reasons. First, if one looks at the historical growth curve of collaboration from 1800 to date, it now displays the unmistakable signs of a logistic curve! What appeared only to be a slight possibility in 1987 has become clearly the most probable outcome, with the addition of new data points from 1992, 1997, and 2002 to the earlier data. All encouragements to collaborate to the contrary, the single author has finally demonstrated an irreducible level of participation at something like one in ten papers overall, with slightly fewer in chemistry (about one in

Proceedings of ISSI 2005, pages 727–730.
Edited by P. Ingwersen and B. Larsen.

sixteen) and slightly more in biology (about one in seven). It is not surprising that the curve is logistic. One might expect a logistic on general grounds, because it is so typical of growth curves in general.

Second, it is also the case that the singly-authored contribution continues to exercise a certain fascination and claim a somewhat unique degree of respect. It clearly is (or unthinkingly is taken to be) the result of just one person, one researcher whose strength of intellect and skill alone is testified to by the publication, against competition, of a paper as worthy as any created by a team of other researchers. There's almost a feeling amongst some researchers that it's good to be able to publish a singly authored paper from time to time – it shows that one still has the right stuff. What's more, to publish alone is also risky, because there's no uncertainty about the allocation of credit or criticism, there's no one else to absorb any negative reactions. In terms of the individualistic history of science, the pantheon of great scientists of the past, one's candidacy for admission would seem to require that some significant fraction of one's research be unequivocally one's own. Consequently, it is no surprise that a single author paper enjoys a certain cachet.

Third, it continues to reinforce the caveat that the apparent practice of collaboration, as measured by sampling a few select journals will almost always produce biased results. Because most journal samples consist of the "core" journals of a field, the ones with the greatest impact factors, that bias usually overestimates the degree of collaboration.

The situation is perhaps least troubling in chemistry, because the single author in that scientific field is indeed practically absent. It's seemingly almost impossible to overestimate the degree of collaboration in chemistry. In a count taken from Chemical Abstracts for 2002, single authors represented about 1/16 authors, or 6%. That is such a small fraction as to amount to little more than "noise" when faced with a "signal" of 15/16 or a 94% frequency of collaboration. Nonetheless, it may represent an irreducible minimum. Even in the most collaborative research field, the individualistic author-researcher still lives, and from time to time, finds a worthy outlet for his or her research.

However, if one turns to the most eminent, reputable, "core" journals in chemistry, it turns out that approximately 100% of their articles are collaboratively authored. To a first approximation, that means that somewhere in the chemical journal hierarchy, there must be some journals whose collaborative frequency must be around 86% (or even less), to balance the excesses of the core journals. Naturally, the distribution of collaborative frequencies could be distributed in such a way that a mass of journals with not much less than 94% collaboration could swamp the core, while still leaving the field as a whole fairly uniform in collaborative frequency, with a very narrow range from 92 to 100%.

The situation is a little clearer in Biology, where the average collaborative frequency in Biology Abstracts for 2002 is 87.83 The greater variation here makes it clearer that the single author still has a significant place in the annals of research publications.

Physics remains the least collaborative of all, although it leads in team research and papers with huge numbers of authors. In 2002, Physics Abstracts displayed a collaborative frequency of 86.02%, meaning that about 1 in 7 papers came from researchers working alone, or at least contributing research results without formally acknowledging any other scientists as co-authors.

The overall average for 2002 of about 12% single authorship represents an almost static situation, compared with 1997, when the count averaged across Biological, Chemical and Physical Abstracts reached approximately 12.5-13%.

However one chooses to interpret the results, it is clear that the nearly universal practice of choosing "core" journals with the greatest impact factors as a sample from which to determine the statistics of multiple authorship always produces biased results. All such results must be interpreted with care.

Teamwork Continues to Expand at the Expense of Collaboration

Let us begin by differentiating the terms collaboration and teamwork according to the number of authors involved, and how those authorships are represented in the statistics of multiply-authored research:

Defn. Collaboration: work which results from about 2-5 co-authors.

Defn. Teamwork: work which results from about 6 or more authors.

From 1800 to 1950, the statistics of co-authorships followed a Poisson distribution. Co-authorships in the social sciences still predominantly follow the Poisson pattern. Since the emergence of "Big Science" after World War II, and the appearance of 5 and 6 authored papers, the statistics of the distribution of authorships in collaboration have typically followed a negative binomial probability distribution. That is, what intuitively and historically we think of as collaboration takes place predominantly among a small number of individuals, who are more or less well known to each other. However, first in physics, and then in chemistry, more and more "teamwork" began to appear.
The following is typical of biology shortly after World War II:

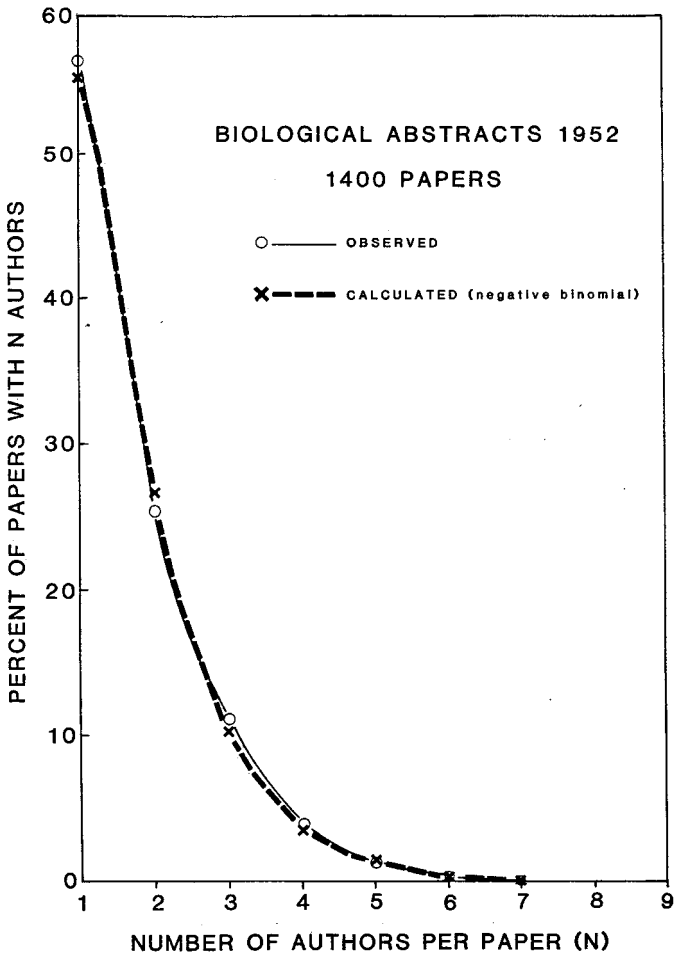

Figure 2. Biology in 1952.

For science as a whole by 1972, the mode in co-authorships was becoming two, and that the negative binomial distribution had just about reached its limit.

On the other hand the statistics of the distribution of co-authorships in teamwork have typically followed an inverse power law probability distribution, $1/n^k$, with k typically between 4 and 6. That is, teamwork, like that which reduces and represents the data from High Energy Physics (HEP) projects or from the Human Genome Project (HUGO), typically involves large numbers of individuals brought together who necessarily develop social relationships in a different way and to a different degree, than can 5 or fewer individuals. Such a radically different type of organization of research is in part a response to radically new (and costly) methods.

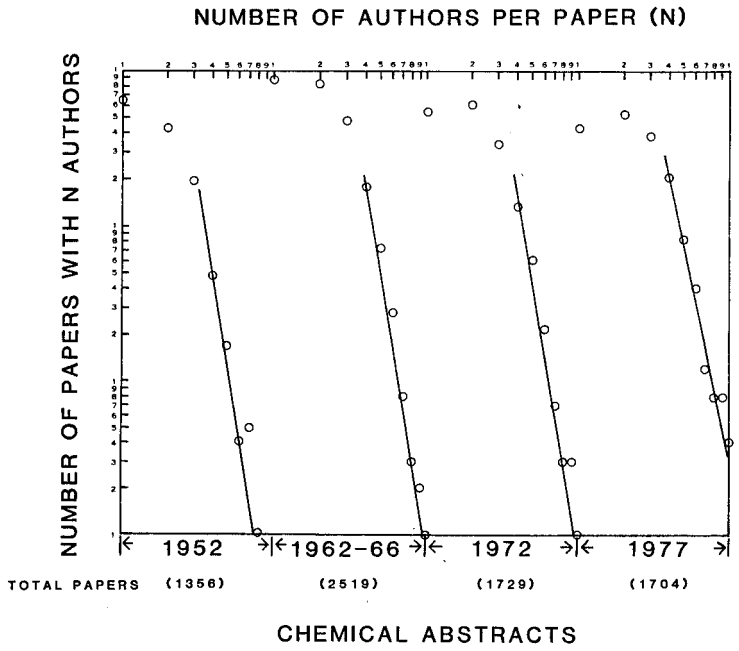

Figure 3. Teamwork graphically displayed. The distribution of co-authorships for multiply-authored papers (6 or more authors) follows an inverse power law distribution.

The most recent counts of the abstracting journals referred to above has enabled more up-to-date representation of these statistics and show that the situation has remained essentially unchanged, with the exception of a growing percentage of papers published with more than 5 authors. In particular, if one traces the development of biology through 1997, one can see the advent of team research at last beginning to affect the distribution of co-authorships from 1987. In all three broad "fields" of the Abstracts, the percentages of papers with 6 or more authors (teamwork) has been growing remarkably in the last 20 years.

The levelling off of the growth of collaboration compared with the single author at around 10% could be taken to suggest a "natural" limit to policy makers favoring collaborative endeavors. Pushing collaboration any further may simply not be worth the effort it might take to disrupt already established patterns research and research organization. Instead, it might be time to revisit the single author's productivity.

To date, no single probability distribution has been found which describes both ends of the multiple author spectrum, the low end of "collaboration," and the high end of "teamwork." The failure of most common probability distributions, and linear combinations of them as joint probability distributions testifies to the robustness of the two categories defined here, of two different types of the organization of research involving more than one individual. Such different types of organizing research suggest possibly characteristic differences in method, in criteria of validity, reliability, and in epistemic authority.