Comments on "Collaboration, Stars and the Changing Organization of Science"

This is a much improved chapter. The authors report some interesting results that are consistent with other work in the book. Most interestingly, the geographic distance between coauthors has increased substantially, notably that the concentration of publications within an institution has decreased and that the institutional rank distance between coauthors has increased. They find that the concentration of publications at the individual level has increased. They also note the pool of potential coauthors has increased. The authors posit that these trends are the result of two factors: the burden of knowledge and collaboration supporting technologies.

I still have a number of unanswered questions, however, which I detail below

1. The context

The authors have chosen evolutionary biology as the field on which to focus. It would help enormously to be provided some context about the nature of the field which might inform the need for collaboration, particularly given the research questions.

a. How collaborative is the field? Is it "big science", like astrophysics, or smaller scale, like chemistry. Has the nature of the production of the science changed (like chemistry).

b. How technology intensive is it? Does it require the use of large scale complex equipment, and what has happened to the price of that equipment over time?

c. How has the nature of the field changed? Does evolutionary biology still mean the same thing now as it did 29 years ago? Or has it now more interdisciplinary and been influenced by the convergence of physics/chemistry/biology?

d. How have other exogenous factors come into play? Have there been big interdisciplinary funding initiatives: funding for nanotechnology and the human genome sequence which have fundamentally changed both the nature and scale of scientific endeavor in a number of fields. How did the doubling of funding from NIH affect this field? Is there any chance of using a sharp change in these factors to identify some of the ideas you posit?

Since graduate students and postdocs are such an important part of the story of potential collaborators, it would seem sensible to set out some basic facts. How important are postdocs and graduate students in the production of this type of science? In other research, I have found (with Paula Stephan and Jacques Mairesse) that the composition and size of the team varies dramatically across research fields (see below). How is this field characterized? How has that changed over time?

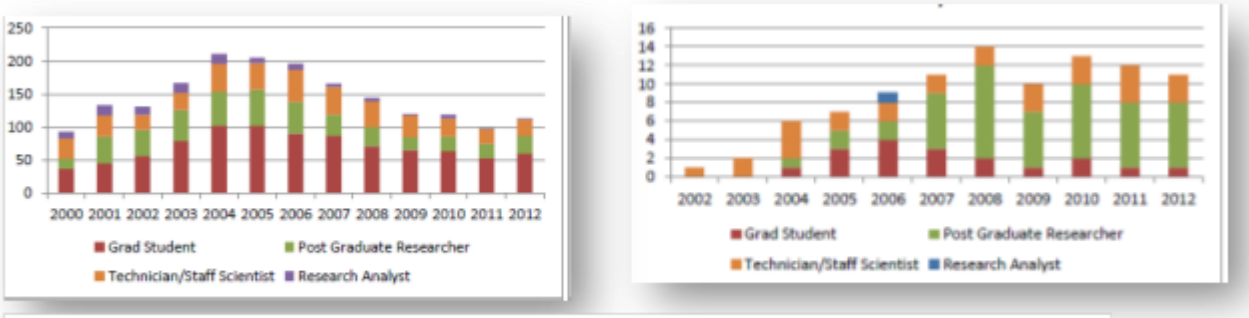

Figure 1: The Composition of Research Teams for Two Different PIs

In a similar vein, what is the market like for postdocs and graduate students? DO they primarily go to industry or academia? How has that proportion changed over time?

After doing the background analysis, the authors might want to conduct some carefully structured interviews with researchers to get an understanding of how the production of science has changed over time

2. Data Questions

I am still uneasy about the structure of the dataset, which seems to be driven by the idiosyncratic imperatives of the Web of Science, and I wonder how much of their findings are due to this structure. The paper is partly about authors and partly about institutions, yet the frame is based on a very specific selection of publications which is the link asset that includes information about both. It would seem to be highly unusual, given the known skewness of productivity of scientists, of institutions, and of the salience of publications, that the selection decisions are representative of all three. The authors should at least provide a discussion of how the selection decisions affect their analysis.

The salient features are below, together with suggestions about how to generalize them

The link asset: publications

a. The frame is first derived from publications in four journals over 29 years whose focus is evolutionary biology.

   It would seem to be appropriate to establish whether the frame has remained constant over time, both with respect to coverage and content. In particular, have any new journals emerged that might siphon off contributors (as has happened in Economics, for example)? There are 15, 256 articles .. is the number per year roughly constant over time, both in levels, and as a proportion of all journals in which evolutionary biologists publish? Your interviews could help identify a list of the key journals, and help provide some of this context.

   It would also seem appropriate to determine what is missing. How many famous evolutionary biology papers are NOT published in these journals, and has this changed over time? Your interviews could help determine this dimension.

b. This frame is then used to collect all articles referenced at least once, and identifies 149,497 articles, which is the "corpus of influence". Is this number per year roughly constant over time? What are these journals? It would help to provide a list of the twenty or so most frequently occurring journals, and see if they do indeed consist of the most "important" journals for evolutionary biology. Your interviews could include presenting the list of journals to your respondents and asking whether the "corpus of influence" makes sense, or whether some other journals should have been included

c. The use of citations weights worries me, and it seems to me the answers to the posed research question should be driven by measures of publications, not by citation weights. If you use citation weights, you are, by definition, heavily weighting the research of stars (who are defined by citations) ..so the whole argument seems very circular. To make an

analogy with research on firms, if the focus of the research is firms, then use firms as the unit of analysis. If the focus of research is employment, use employment weighted firms. In the former case, the small startup has equal weighting as GM. In the latter case, the paper will be primarily about GM's behavior. At the very least ,the paper should discuss the implications of this choice, rather than devote a single line at the end of page 5

The authors

d. There are 171,428 authors thus identified. 140,240 of these are dropped because there are no more than two publications linked to their name. I think the decision to drop more than 80% of the sample requires some more discussion, because it affects key measures.
   i. Distance measures. How many of these are dropped because it is not possible to identify non unique author names (ie, for example, Asian authors)?
   ii. Time variability. How does the sample selection change over time?
   iii. Institutional variation. How many institutions are dropped?
e. Fractionally weighted output measures. The decision to weight author output measures by the full publications, given that coauthorship has increased over time, effectively weights output in later time periods more heavily than previous time periods. This decision substantively affects the econometric results. At the least, both weighted and unweighted results should be presented and discussed throughout.
f. Relationship among authors. The authors make much of the post doc/graduate student mentorship relationship later in the paper. Is there any information that can be gleaned from the dataset to substantiate some of the key hypotheses? Can a subset of the papers be pulled and the relationships confirmed?

The institutions

g. Institutional locations. Just over half (57%) of the papers have a single institution listed, and so all authors are located with this institution, and 79% of authors are attributed to an institution. It would be very helpful to know how many institutions are at risk to being included, and how many (and which ones) are dropped.
   i. Temporally variability How does this change over time?
   ii. Location: How many are US and how many are overseas (particularly in Asia)
   iii. Coauthorship: How does the selection decision change the number of coauthors included in the dataset?

3. Model

The authors posit that the activity is due to an increasing pool of post docs, but offer little evidence for this speculation. The model is really interesting, but might be better applied to a different paper, since it is all about graduate students, and there is no data to test the model.

There are at least three alternative hypotheses, which could be helped by the interviews, and could potentially be modelled.

a. Increasing focus on funding agencies. Do "distant" collaboration explicitly cite funding more frequently than "near" collaborations

b. Burden of knowledge. Our interviews with researchers suggest that they go out and look for experts in particular areas to collaborate with, because they don't have the knowledge to apply a particular scientific technique. Do "distant" collaborations include coauthors who have published in areas with specific technical requirements?

c. There is a lot of discussion of the effect of collocation, which would be helped by a discussion of evidence of collocation within the data.

I am not sure that these can be tested, so maybe the authors should just say that they have uncovered some interesting results, and leave the chapter as a descriptive exercise.

4. Generalizability

As noted above, the results are interesting, and very congruent with the other chapters in the book. However, it would be useful to close with a broader discussion fo the generalizability to other areas of science.

Julia Lane
American Institutes for Research