Scientific Superstars and their Effect on the Evolution of Science

Richard Klavans and Kevin W. Boyack

(rklavans, kboyack)@mapofscience.com
SciTech Strategies, Inc.

Conference theme: Innovation indicators

Introduction

This study bridges two areas of research in bibliometrics. The first focuses on the well-established fact that a very few authors are extremely prolific [1]. We call these individuals 'superstars', and utilize data from Scopus to provide the first large scale study of the publication behavior of this unique segment of the scientific population.

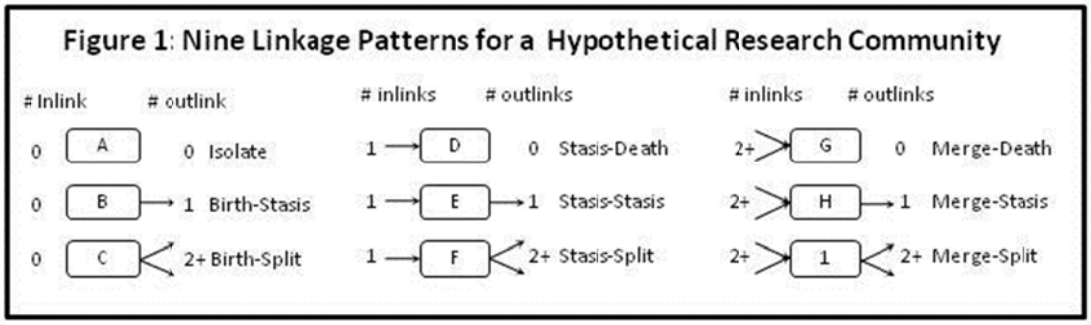
The second area of research focuses on changes in the structure of science. We build on a long tradition of using co-citation analysis to model the socio-cognitive structure of research [2], and the more recent studies that suggest that these research communities might split apart and merge over time [3, 4]. Specifically, we use a database that classifies scientific documents into relatively small research communities (approximately 15 papers per community per year) and tracks the linkages between these research communities over time. This database suggests that there is a great deal of birth and immediate mortality in socio-cognitive structure. 32.8% of the communities (containing 25% of the articles) are coded as *isolates* (research communities having no significant linkage to prior or subsequent research communities).

Our hypotheses are relatively straightforward. First, we test whether other scientists follow superstars. If this is true, research communities with a superstar will be more likely to survive, and those without leaders will be more likely to die; superstars will be less likely to publish in isolates. Second, we test whether superstars are prescient. If this is true, they will avoid publishing in research communities that will not survive to the subsequent year. And finally, we test whether superstars are disruptive – that is they tend to avoid situations of stasis (research communities with only one input and one output link). Following is a description of the data, hypotheses and a short summary of our major findings. We provide evidence that superstars have a unique impact on how science evolves by being leaders, prescient and disruptive.

Data

We used the 2000-2008 Scopus database (as of November 2009), consisting of 10.36 million scientific articles published in 554 disciplines by 8.86 million authors. Articles were partitioned into research communities by year using co-citation analysis [3]. Each community is assigned to a discipline using article-journal distributions [5]. Communities are linked from year to year using reference overlapping. This method differs from the current practice of using multi-year sampling and significantly higher levels of aggregation [6-8]. Multi-year sampling and larger levels of aggregation reduces structural instability; the resulting networks have very few isolates and much large networks of research communities. We prefer to highlight this instability as evidence that there is a significant amount of socio-cognitive experimentation in science. The fact that many experiments in science fail is an important aspect of the phenomenon, and critical to gaining an understanding of how these structures evolve. Our method highlights this instability. In addition, the lower levels of aggregation are more consistent with early observations of the approximate size of a research community [9].

Since we are only considering links to prior and subsequent years, our analysis focused on the research communities from 2001 to 2007 and their links to prior and subsequent years. Figure 1 illustrates the nine possible linkage relationships for a research community. Each research community could have three possible inputs (no links, one link and 2+links) and the same three possible outputs. A research community that has no input and output links is called an *isolate*. All research communities with no input links but positive output links are examples of *birth*. *Stasis* is the situation where there is only one input and output link. Splitting occurs when the output link has two (or more) links. Merging occurs when the input link has two (or more) links.



Scopus data include unique author ids. Although there are some authors with multiple author ids, and some few ids merging authors, it is the best such system available that is applied to all of science. To identify superstars, we first calculated the dominant discipline for each author. Data were then limited to the 236 disciplines with at least 10,000 authors. Superstars were identified as the top 1% of these authors by discipline, ranked by publication level, resulting in a set of over 75,000 superstar authors. Publication patterns of these superstar authors by linkage type were compared to those from the remaining 99% of authors, by discipline.

Hypotheses

H1: Leadership: $S_a/S_{a>i} < D_a/D_{a>i}$. The percentage of papers in isolate research communities by superstars [$S_a/S_{a>i}$] will be less than the percentage of papers in isolate research communities by all authors in that discipline [$D_a/D_{a>i}$]. Our reasoning for this hypothesis is that superstars will publish less in isolates because followers will tend to follow the superstars, and in doing so, will increase the likelihood that the community will persist. The communities without superstars are therefore less likely to survive. Adjustments for disciplinary norms are necessary because there are significant variances in the percentage of isolates by discipline.

H2: Prescience: $S_{(d+g)}/S_{b>i} < D_{(d+g)}/D_{b>i}$. In this case, we assume that the superstar avoids research communities that will die (set D and G). Our reasoning for hypothesis 2 is that a superstar is more likely than the average researcher to notice that a particular research topic isn't going anywhere. Superstars exit before a stream dries up and don't enter streams that are about to dry up. We exclude isolates from this calculation. Note that in H2 (and H3) the sample is conditional (researchers that only published in isolates will not be in this sample).

H3: Disruption: $S_{(f+h+i)}/S_{(e+f+h+i)} < D_{(f+h+i)}/D_{(e+f+h+i)}$. Our reasoning for hypothesis 3 is that superstars are actively interested in making a significant impact on science. They will therefore tend to select research communities that are dynamic (where there is merging and splitting), and avoid those areas where the status quo is being maintained. Researchers whose publications were totally in (the union of) set A, B, C, D and G will not be in this sample.

Findings:

All three hypotheses are strongly supported by the data (see Table 1). 78.1% of the superstars published less than the expected amount in isolates. 81.4% of the superstars published less than the expected amount in research communities that were dying branches. 78.9% of the superstars published less than the expected amount in stasis research communities. (Expected values are 50% in all three cases).

Table 1. Cross tabulations for hypotheses 1, 2 and 3, reporting numbers of authors. Sample sizes are not equivalent due to conditional sampling in the hypotheses.

	Hypothesis 1		Hypothesis 2		Hypothesis 3	
	No	Yes	No	Yes	No	Yes
Superstars	16,500	58,954	13,921	61,104	15,453	57,761
Remaining 99%	2,947,678	2,905,103	2,382,536	2,335,483	1,711,360	1,669,162

For each author, we limited the analysis to research communities in the discipline that the author was assigned to. We also excluded any research communities containing less than 1% of the output of the author. This can occur because papers are fractionally assigned to research communities using reference matching. For example, it is possible for 1% of a paper to be assigned to a research community if the research community has 100 unique references and the citing paper only has one reference in common.

There are many additional findings emerging from this analysis. For example, our initial analysis suggests that these effects are true throughout the entire publication range (i.e. not limited to the top 1%). The data also suggest different publication strategies that superstars pursue that could be the basis for innovation indicators. Space does not permit further elaboration of our findings and implications for further research.

References:

1. Price, D.J.D., *Little Science, Big Science*. 1963, New York: Columbia University Press.
2. Small, H., *Co-citation in the scientific literature: A new measure of the relationship between two documents.* Journal of the American Society for Information Science, 1973. **24**: p. 265-269.
3. Klavans, R. and K.W. Boyack, *Using global mapping to create more accurate document-level maps of research fields.* Journal of the American Society for Information Science and Technology, 2011. **62**(1): p. 1-18.
4. Palla, G., A.-L. Barabási, and T. Vicsek, *Quantifying social group evolution.* Nature, 2007. **446**: p. 664-667.
5. Klavans, R. and K.W. Boyack, *Toward an objective, reliable and accurate method for measuring research leadership.* Scientometrics, 2010. **82**(3): p. 539-553.
6. Chen, C., F. Ibekwe-SanJuan, and J. Hou, *The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis.* Journal of the American Society for Information Science and Technology, 2010. **61**(7): p. 1386-1409.
7. Upham, S.P. and H. Small, *Emerging research fronts in science and technology: Patterns of new knowledge development.* Scientometrics, 2010. **83**(1): p. 15-38.
8. Upham, S.P., L. Rosenkopf, and L.H. Ungar, *Innovating knowledge communities: An analysis of group collaboration and competition in science and technology.* Scientometrics, 2010. **83**(2): p. 525-554.
9. Franklin, J.J. and R. Johnston, *Co-citation bibliometric modeling as a tool for S&T policy and R&D management: Issues, applications, and developments*, in *Handbook of Quantitative Studies of Science and Technology*, A.F.J. van Raan, Editor. 1988, Elsevier Science Publishers, B.V.: North-Holland. p. 325-389.