Understanding Institutional Collaboration networks: Computer Science vs. Psychology

[1]Jiadi Yao, Les Carr[1] and Stevan Harnad[2]

[1] *jy2e08@ecs.soton.ac.uk*
School of Electronic and Computer Science, University of Southampton, UK

[2] Department of Psychology, Université du Québec à Montréal, Canada

Abstract

Institutions assume that if they are more productive (i.e., publish more papers), they will produce more high quality research. They also assume that if they collaborate more, they will be more productive. We test these causal assumptions using nearly 30 years of worldwide publication and citation data in Computer Science and Psychology. Four quality metrics, three collaboration metrics and one productivity metric were used. Spearman's Rank Order non-parametric correlation shows that these three groups of variables are highly inter-correlated. Regression analysis was used to partial out the effect of the third variable and reveal the independent correlation between each pair of the variables.

In Computer Science, the more productive institutions publish higher quality research as measured by citation counts (including citation counts recursively weighted by the citation counts of the citing institution); the effect is the same, but not as strong, in Psychology. Higher average paper quality in both Computer Science and Psychology are more likely to be a result of greater institutional collaboration than of higher institutional productivity. The proportion of the institutional collaboration is closely linked to institutional quality and productivity. The more proportionally collaborated institutions in fact are less qualitative as well as less productive.

Introduction

Institutions assume that if they are more productive (i.e., publish more papers), they will produce more high quality research. They also assume that if they collaborate more, they will be more productive. The trend for collaborative research has steadily increased over recent decades (Beaver and Rosen, 1978`; Newman, 2004; Choi, 2012), and research has shown the benefits of the collaboration (Katz and Hicks, 1997; Katz and Martin, 1997; Lariviere *et al.*, 2006; Almendral *et al.*, 2007). European Union research policies support collaboration, encourage creation of institutional networks, sharing of knowledge and promoting innovation. Research programs -- for example, the Framework Programme (FP) -- are established to fund research across the member states, encouraging cross institutional collaboration.

In this paper, we exam some causal assumptions about the relations among productivity, quality and collaboration using nearly 30 years of worldwide publication and citation data in Computer Science and Psychology.

As early as 1926, Lotka (1926) studied the relationship between the number of individuals at different productivity levels: the number of researchers who publish one paper per year is two orders of magnitude more than the number who publish 10 papers, and four orders of magnitude more than the number who publish 100 papers. This is referred to as Lotka's Law of productivity. Price and Beaver (1966) showed that the number of collaborators is positively correlated with the number of articles published by the author. Through qualitative analysis, they also found that the most prolific researchers also collaborate most. A year later, Zuckerman (1967) interviewed 41 Nobel laureates in science disciplines and identified a strong relationship between collaboration and productivity. Laureates published more papers and also collaborated

more than a matched sample of scientists. Pravdić and Oluić-Vuković (1986), using collaboration and productivity data in Chemistry, found that research output is correlated with frequency of collaboration. After interviewing a sample of the authors, they also learned that collaborating with high productivity authors is positively correlated with personal productivity whereas collaborating with low productivity is negatively correlated. Glänzel and Schubert (2004) consider collaborations at three different levels: person level co-authorship, cross-country co-authorship and multi-country co-authorship. For all three levels, co-authorship is positively correlated with productivity.
The same positive correlation is found by Adams *et al.* (2005) between the size of collaboration groups and the scientific productivity.

Lee and Bozeman appear to have found something subtler. In their 2003 research report (Bozeman and Lee, 2003), they used a regression model to determine whether the predictive power of collaboration depended on factors such as job satisfaction, rank, age, gender etc. They surveyed and interviewed 443 academics to obtain their data and then calculated the regression. Regardless of any further variables added to the regression, the number of collaborators of an author remained the strongest predictor of productivity. However, in their later paper on the same topic (Lee and Bozeman, 2005), they extended the article and book counting method in two ways: fractional counts (each author gets an equal fraction of the credits for collaborative papers) and full counts. While number of journal papers was still strongly and significantly correlated with number of collaborators, there was no significant correlation between number of collaborators and publication counts when using fractional counts. It seems that different counting methodology can potentially lead to very different results, however, the exact counting methodology is often omitted in the literature.

More recently, Defazio *et al.* (2009), using the EU framework programme to study these variables in Chemistry, found that researchers tend to collaborate just to secure funding, the impact of funding on productivity is positive, but the impact of collaboration is weak. By splitting the period into pre-funding, during-funding and post-funding periods, they found that collaboration during funding does not correlate with productivity; post-funding, however, although collaboration decreases, it has a strong positive correlation with productivity. So it appears that the connections the researchers established pre-funding and during-funding went on to have a positive effect on subsequent research output.

The research discussed so far was all based on cross sectional data, making cause-effect inferences untestable. He *et al.* (2009) constructed a longitudinal dataset of 65 New Zealand researchers for 14 years. Among other findings, they claimed that international collaborations are positively related with future research output. However, they could not find any significant correlation with future output for within-university collaboration and domestic collaboration.

Data and Method

Data Source and Descriptive Statistics

The source data are from Thomson-Reuters Web of Science, covering Computer Science and Psychology papers published from 1973 to 2010. The papers are stored in a MySQL database in the format of four separate tables describing the article's subject, author, institution and citation counts. There are 479,913 Computer Science papers, of which 164,553 (34%) are multi-institution collaborative papers, 277,425 (58%) are single-institution papers and the remaining 37,935 (8%) do not have any institution specified. 267,666 papers have at least 1 citation, with

an overall nonzero citation rate of 56%. The total number of citations received by the Computer Science papers was 2,711,196.

There are 208,066 Psychology papers, covering General Psychology, Clinical Psychology and Social Psychology, hence less than half the number of Computer Science papers published in the same period. Of these 68,141 (33%) papers are multi-institution collaborative papers, 130,891 (63%) single-institution papers, while the remaining 9,034 (4%) do not have institutional information. The percentage distribution of institutionally collaborated papers is very close to that of Computer Science. Psychology is a more highly cited discipline than Computer Science. Of all the Psychology papers, 156,992 received at least 1 citation, this accounts for 75% of total papers -- a much higher percentage than in Computer Science. Total citations for Psychology papers were 3,514,787: almost 1 million citations more than Computer Science on less than half the number of papers.

1,125 institutions published more than 100 Computer Science papers in the period of 1973 to 2010. Of these 8% are companies, 5% are research institutes, while the remaining 88% are universities. Of the 88% universities, 698 (71%) universities are identifiable through a matching name in Webometrics university ranking. Psychology is a less commercially applied discipline than Computer Science: no company is listed as the institutional affiliation for Psychology papers. Out of 542 institutions, 10% are research institutions, while 90% are universities. Among these universities, our algorithm was able to identify 414 (88%) universities. The analysis presented in this paper is based only on the identified universities.

Description of Metrics

Correlations were analysed for three institutional variables: collaboration, productivity and quality using the following metrics:

Productivity (P) 1973-2010 was measured by total institutional papers output
Collaboration (C) 1973-2010 was measured by:

- Number of Collaborative Papers (CN)
- Size-weighted Collaboration(CS)
- Percentage Collaboration (CP)

CN, the total number of papers with at least two distinct institutional affiliations, is the number of *cross institutional collaborative* papers an institution has published according to WOS 1973-2010. CS is CN weighted by the size of the collaboration: Instead of treating every collaborative paper equally, papers with more authors contributing and more institutions participating are assigned a higher score. The formula to calculate CS is:

$$CS_i = \sum_p A_{pi} \times (TA_p - A_{pi})$$

Where A_{pi} is the number of authors from institution i on paper p, TA_p is the total number of authors for paper p.

CP is the ratio of an institutions papers that are collaborative: the ratio of C to P, for institution i:

$$CP_i = \frac{C_i}{P_i}$$

Quality (Q) is estimated with four metrics, three citation based, 1973-2010, and one institutional ranking based:

- Citations per institution (QC)
- PageRanked citations per institution (QPR) (incoming citations weighted by the citation weight of the citing institution)
- Citations Per Paper (QCP)
- Institutional Webometrics Rank (QW), July-2010 version.

The abbreviations for quality variables all start with Q. QC is the sum of all the citations received by the papers published by the institution 1973-2010. It measures overall institutional impact and quality. QPR is derived by applying PageRank algorithm. QPR recursively weights the cited institution's citation count by the citing institution's citation count. Both QC and QPR are measures of institutional quality in a given discipline. QCP is the institution's average citation count per paper, calculated by dividing institution's total citation QC by its total paper output P. This becomes an institutional size-normalised quality metric, or simply put, the institution's average paper quality. This quality metric is, in general, closer to what institutions hope to increase with higher productivity. QW is the rank of the institution according to the July 2010 version of the Webometrics ranking. The Webometrics rank is itself a composite metric derived from some of the other quality, collaboration and productivity metrics used in this study. Hence QW is not an independent variable: a strong correlation is expected with productivity. To make the result easier to interpret, the rank ordering is inverted, so high rank indicates higher quality; this way, a positive correlation corresponds to positive relationships between quality and the other two variables.

In the remaining sections, bold and italic variable names represent the raw, and un-filtered state, while italics-only represents the variable with the other effects partialled out. For example, ***P*** represents the original paper count for institutions, while *P* represents the paper count with either collaboration or quality effects removed.

Method

Spearman's Rank Order correlation rho (ρ) is used to measure the associations between the pairs of variables. The value of ρ varies between -1 to 1, where -1 is a perfect negative correlation and 1 is a positive correlation. 0 means no correlation. No transformation of the original data is needed because this non-parametric metric makes no assumptions about distribution. Multiple regression was used to partial out the effect of the third variable:

$$y = \beta \boldsymbol{X} + \epsilon$$

The dependent variable y is expressed as a vector of independent variables **X** with a linear transformation β, plus the residual ϵ, which is the unpredicted portion of the variance. This is the portion of y left after removing **X**. In order to apply ordinary partial correlation, both the dependent variable y and the independent variable **X** would have had to be normally distributed. However, the distribution of publication number, collaboration metrics and citations is generally not normal(Barabási and Albert, 1999; Newman, 2001). Instead, they are distributed according to a power law, with most institutions publishing a small number of papers and most papers published by a few institutions. A power transformation is needed to convert these skewed distributions into a normal distribution before applying the partial correlation.

To determine the best λ value for the power transformation $y = x^{\lambda}$, the Box-Cox technique implemented in SPSS was applied. Box-Cox tests a series of λ values and plots the distribution

of the transformed variable. Q-Q plot histograms are drawn by the algorithm for manual selection of the transformation best approximating normality.

Results and Discussion

Productivity and Quality

The institutional productivity (***P***) and institutional disciplinary research quality variables (***QC*** and ***QPR***) show statistically significant, positive and high correlations in both disciplines (Figure.1). These two pairs of correlation coefficients (***P*** with ***QC*** and ***P*** with ***QPR***) in both disciplines reach the high 0.8 range. When the collaboration effects indexed by institutional collaborative papers (***CN***), size-weighted collaboration (***CS***) and percentage collaboration (***CP***) are removed one by one completely from ***P*** and ***QPR*** using partial correlation, the correlation coefficient is reduced substantially in both disciplines, but is still statistically significant and positive. In Psychology, the correlation between *P* and *QC* dropped to 0.364 from 0.879 after the removal of collaboration effects, while the correlation between *P* and *QPR* dropped to 0.056 from 0.856. Computer Science had a similar major reduction after partialling out collaboration, from 0.905 to 0.438 between *P* and *QC*, and from 0.885 to 0.276 between *P* and *QPR*.

In both disciplines average paper quality (***QCP***) per institution has a significant medium-sized positive correlation with the number of institutional papers published (***P***): 0.501 for Computer Science and a higher 0.537 for Psychology. However, in Computer Science the correlation disappears once collaboration effects are removed. In Psychology, a large reduction also occurs, but the correlation remains significantly positive.

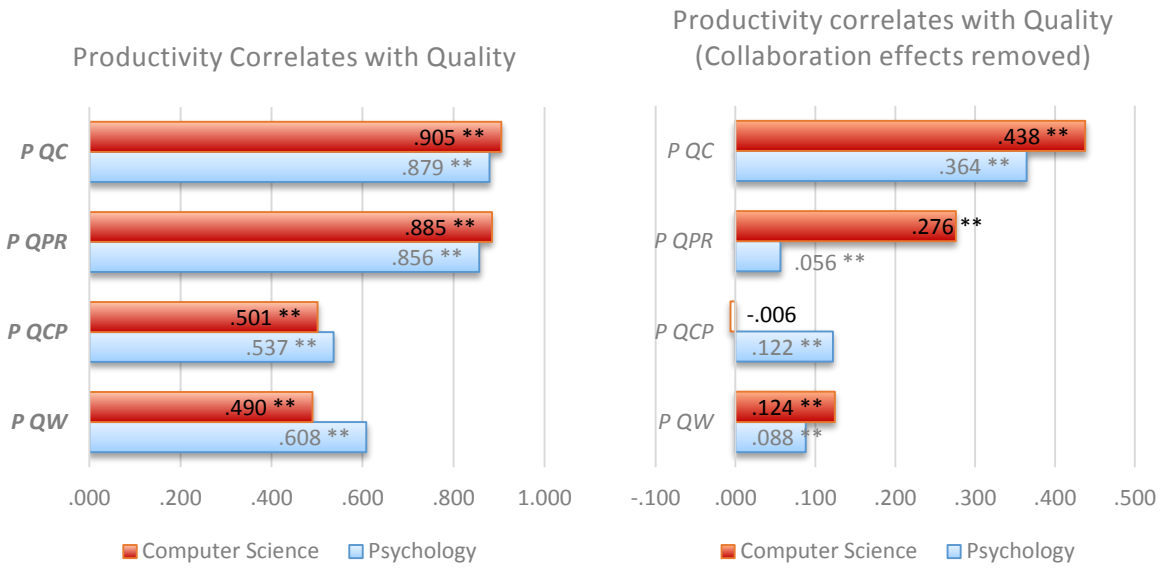

Figure 1. Correlation between productivity P and quality (*QC, QPR, QCP, QW*). Left: pairwise correlations, without removing collaboration effects (*CN, CS, CP*). Right: correlations after partialling out collaboration effects. Large reductions are observed in all pairs of correlations. In both disciplines the correlations between *P* and *QC*, *P* and between *QPR* were significant, positive and high. Psychology has stronger correlations than Computer Science between *P* and *QCP* and between *P* and *QW*. In Computer Science, P and QCP are not correlated once the collaboration effects are removed, whereas in Psychology, the correlation is reduced to 0.122 from 0.537. ** indicates p<0.01

Collaboration emerges as one of the important factors affecting average institutional paper quality. No correlation remains with productivity once collaboration effect has been partialled out. In Computer Science, institutions publishing more papers are not likely to have higher average paper quality, whereas institutions with more collaborative papers may have higher average paper quality. In Psychology, in contrast, institutions publishing more papers are likely to have higher average paper quality. In Psychology, collaboration has the biggest impact on PageRank-weighted citation counts (*QPR*), as removing collaboration almost obliterated the correlation between paper (*P*) and *QPR*.

Collaboration and Productivity

Figure 2 shows in both disciplines that before controlling of the effects of quality, there is a strong positive correlation between productivity (***P***) and collaborative paper counts (***CN***), and between ***P*** and ***CS***. Institutional percentage collaboration (***CP***) has a medium-sized negative correlation with institutional productivity. Removing quality effects did not affect the correlations as strongly as removing collaboration effects, suggesting that the quality effects are more independent of the other two factors.

In Computer Science, controlling quality effects reduces the negative correlation but it is still statistically significant. In contrast, in Psychology the negative correlation increases. Institutions' percentage collaboration is negatively correlation with their number of papers, both before and after quality effects are removed, suggesting that as institutions become more productive, they collaborate less. This might be because resources, funding and opportunities for collaboration reach a ceiling effect in the most productive institutions or because the most productive institutions are also those with the most resources and elite researchers who can easily conduct research on their own, without needed of collaborators. At the high end, the costs may exceed the benefits of collaboration. This suggests that degree of collaboration alone is not a major factor in determining institutional quality. Institutions who have tried to collaborate as much as possible, enjoining their researchers to publish as many collaborative papers as they could, do not achieve higher productivity as a result.

Collaboration and Quality

Figure 3 shows the correlation between collaboration and quality. The pairwise correlations between the untransformed collaboration and quality metrics, i.e., between CN and QC, CN and QPR, CS and QC, and CS and QPR have high and positive correlation in both disciplines. In Computer Science, the percentage collaboration is correlated negatively with all 4 quality variables. Psychology has a smaller: 3 of the 4 quality variables are statistically significant; there is no correlation between citations per paper (QCP) and percentage of collaboration (CP).

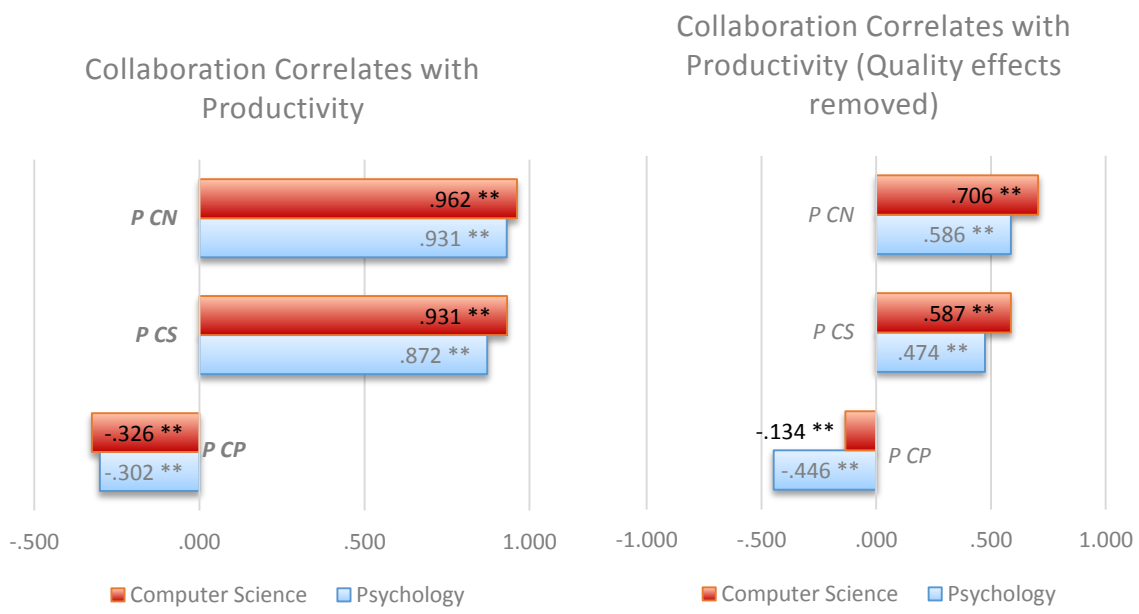

Figure 2. Correlation between collaboration and productivity. Left: pairwise correlations before removing quality; Right: after partialling out quality. Both the *P CN* and *P CS* correlations drop after quality effects are removed, but the reduction is not as great as when collaboration effects are removed. ** indicates p<0.01

The partialling out of institutional productivity adjusts the collaboration and quality variables according to output size, thereby increasing the lower productive institution's collaboration and quality variables accordingly. In Computer Science, the partialling largely removed the relationship of raw citation counts (*QC)* and PageRank-weighted citation counts (*QPR*) with any of the collaboration variables, except the correlation between citation *QPR* and collaboration (*CS*). Therefore, institutional quality and impact as measured by institutional citation counts and PageRank-weighted citation counts is independent of institutional collaboration. Interestingly, the partial correlation did not reduce Psychology's correlation between the same pairs of variables, which remained positive and significant, but small.

The large reduction of the high correlation between *QC* and *CN*, *QC* and *CS*, *QPR* and *CN*, and *QPR* and *CS* when productivity effects were controlled for shows that productivity is the factor linking both collaboration and quality in Computer Science. However, productivity in Psychology is not as strongly affected.

Collaboration is positively with average paper quality both before, and after removing the productivity effect. Psychology has a higher correlation than Computer Science. Partialling out productivity, reduces both correlations, but they remain significant and positive. Institutions that publish more collaborative papers have higher average paper citations counts than those that collaborate less.

Institutional Webometrics ranking is also correlated positively with total collaboration and negatively with percentage collaboration. This agrees with citation based institutional quality variables (***QC, QPR***) in both disciplines. Partialling out productivity changed the correlation between collaborative papers (*CN*) and size-weighted collaboration (*CS*) from positive to negative in Computer Science. In Psychology, the correlations of *QW* with *CN* and *CS* are

significant and positive, but very small. Institutional ranking has no correlation with institutional paper quality in Psychology.

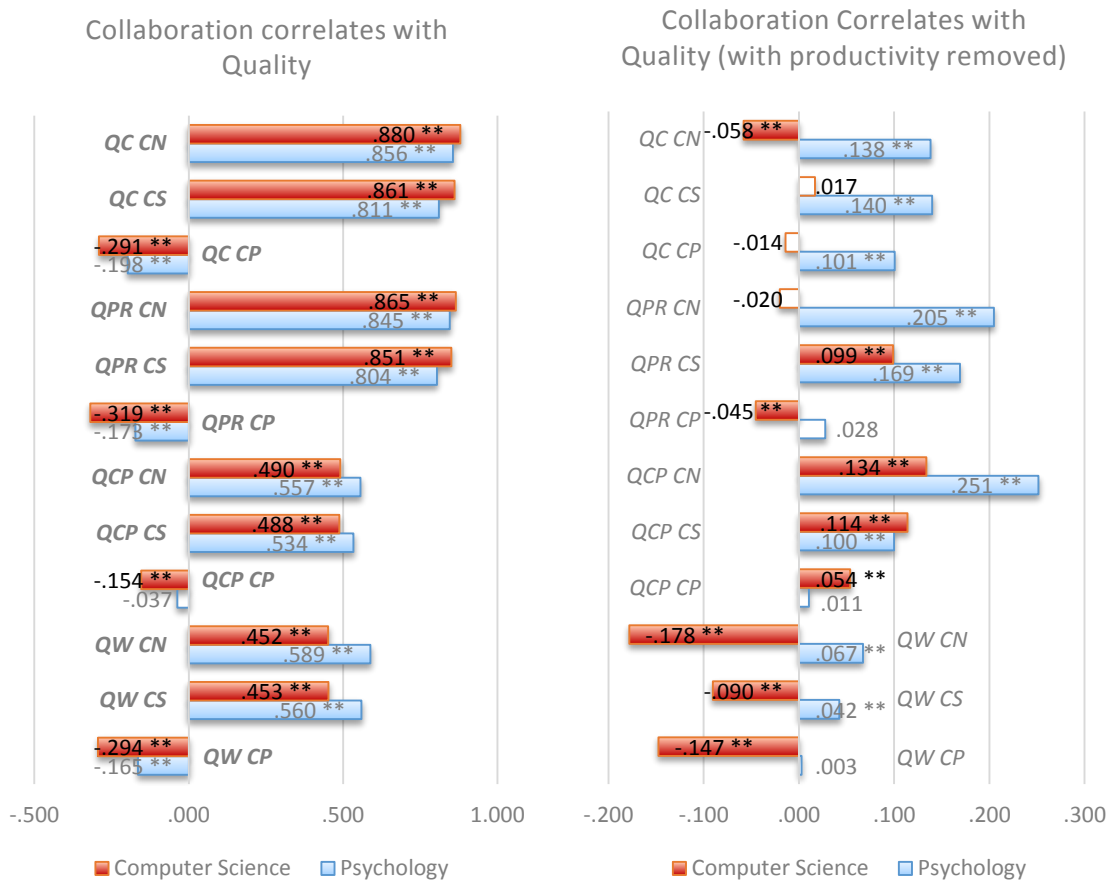

Figure 3. Correlation between quality and collaboration variables. Left: pairwise correlation between original variables; right, correlation AFTER productivity effects removed. There are strong positive correlations between the untransformed pairs of quality and collaboration variables -- *QC CN, QC CS,QPR CN* and *QPR CS* -- whereas the correlations with percentage collaboration in both disciplines are negative. Strong reduction of all correlations is observed after the removal of the productivity effect. ** indicates p<0.01

Conclusion

In Computer Science, the research of institutions with higher publication counts tends to be of higher quality (as measured by total citation counts, both un-weighted and PageRank-weighted); but institutions' citation counts turn out to rise no higher with higher publication counts if the effects of collaboration are removed. Hence more collaborative research publication seems to be one of the factors underlying the higher quality of more productive institutions. In Psychology, with higher publication counts, institutional quality measured by citation counts tends to be higher; but when quality is measured by total PageRank-weighted citation counts – a metric weighted on the basis of how cited the citing institutions are – its correlation with publication productivity is weak.

Collaboration was found to be strongly correlated with productivity, both before and after controlling for quality. Institutions publishing more collaborative papers also have higher total

publication counts. On the other hand, the percentage collaboration (ratio of collaborative papers to total papers) is negatively correlated with productivity variables in both disciplines, before and after the quality variables have been partialled out. This means that institutions that focus more on collaboration than single institution papers do not have a productivity increase. Hence increasing collaboration ratio alone is not a shortcut for increasing productivity. In Computer Science, collaboration was also found to be one of the main factors that affect institutional paper quality. Higher paper quality is a result of more collaborations, rather than more papers published.

Acknowledgement

This research is funded by the Web Science Doctoral Training Centre, University of Southampton. Many thanks to Vincent Lariviere and Yves Gingras for their discussion on the topic and for providing the WoS source data.

Reference

Adams JD, Black GC, Clemmons JR, Stephan PE (2005). Scientific teams and institutional collaborations: Evidence from US universities, 1981--1999. *Research Policy* **34**(3)**:** 259-285.

Almendral JA, Oliveira JG, López L, Mendes JFF, Sanjuán MAF (2007). The network of scientific collaborations within the European framework programme. *Physica A: Statistical Mechanics and its Applications* **384**(2)**:** 675 - 683.

Barabási AL, Albert R (1999). Emergence of scaling in random networks. *Science* **286**(5439)**:** 509.

Beaver Dd, Rosen R (1978). Studies in scientific collaboration. *Scientometrics* **1:** 65-84.

Bozeman B, Lee S (2003). The Impact of Research Collaboration on Scientific Productivity.

Choi S (2012). Core-periphery, new clusters, or rising stars?: international scientific collaboration among 'advanced' countries in the era of globalization. *Scientometrics* **90:** 25-41.

Defazio D, Lockett A, Wright M (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy* **38**(2)**:** 293 - 305.

Glänzel W, Schubert A (2004). Analyzing scientific networks through co-authorship.

He Z-L, Geng X-S, Campbell-Hunt C (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy* **38**(2)**:** 306 - 317.

Katz J, Hicks D (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics* **40:** 541-554.

Katz JS, Martin BR (1997). What is research collaboration? *Research Policy* **26**(1)**:** 1 - 18.

Lariviere V, Gingras Y, Archambault E, ric (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics* **68**(3)**:** 519-533.

Lee S, Bozeman B (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science* **35**(5)**:** pp. 673-702.

Lotka AJ (1926). The Frequency distribution of scientific productivity. *Journal of Washington Academy Sciences* **16:** 317-323.

Newman MEJ (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* **101**(90001)**:** 5200-5205.

Newman MEJ (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* **64**(1)**:** 16131.

Pravdić N, Oluić-Vuković V (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics* **10**(5)**:** 259-280.

Price DJS, Beaver D (1966). Collaboration in an invisible college. *American Psychologist* **21**(11)**:** 1011.

Zuckerman H (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review***:** 391-403.