Collaboration and productivity: an investigation into "Scientometrics" and "UHasselt" repository

Collaboration and productivity: an investigation in Scientometrics and in a university repository

by

L. Egghe[1,2], M. Goovaerts[1] and H. Kretschmer[3,4]

1. Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[5]
2. Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
3. WISELAB, Dalian University of Technology, Dalian, 116024 China
4. COLLNET Center, Borgsdorfer Str. 5, D-16540 Hohen Neuendorf, Germany

leo.egghe@uhasselt.be
marc.goovaerts@uhasselt.be
kretschmer.h@onlinehome.de

ABSTRACT

In this paper we investigate the following problem: for a fixed field or institute, can we prove that, the higher the number of papers of an author (calculated in the total way), the higher

[5] Permanent address

Key words and phrases: collaboration, productivity, co-author
Acknowledgements: The first named author is grateful to Profs. Dr. C. Borgman, J. Furner, W. Glänzel and R. Rousseau for interesting discussions on the topic of this paper. The authors are grateful to Prof. Dr. W. Glänzel for providing the Scientometrics data. Programmers: Walter Brebels and Norbert Gesch. Technical assistance: Theo Kretschmer.

his/her fraction of co-authored papers (with at least one co-author) ? This is one of the possible formulations of the relation between collaboration and production.

We investigate two different data sets: all articles in Scientometrics (from Volume 1 onwards, up to the end of 2007) and the institutional repository of the Universiteit Hasselt (UHasselt), containing 7,604 articles (Feb. 27, 2008). In the Scientometrics case we cannot prove the asserted relation while in the UHasselt case we can prove that high productivity leads to high fractions of co-authored papers (but low productivity can have low or high fractions of co-authored papers).

Since the majority of the papers in the UHasselt repository is in the sciences while Scientometrics is more in the social sciences (certainly in the earlier years) this might be one explanation of the different findings. We also show that collaboration in the case of the UHasselt repository is much higher than in the Scientometrics case.

We find that a determining factor of collaboration is the scientific environment and we present a production collaboration graph on which we can base ourselves for comparisons between fields and/or institutes.

I. Introduction

The relation between productivity and collaboration is an intriguing problem. Intuitively one is inclined to say that higher collaboration leads to higher productivity (the opposite would be surprising !). There are several ways to “explain” such a philosophical assertion: trivially authors involved in co-authored papers have more time to write additional papers since part of the work is done by the other co-authors. Also collaboration could be higher between the “better” researchers which then leads to higher productions. Also collaboration is higher in those fields (as the sciences) where one has (large) research labs yielding a high productivity.

Before we investigate the above mentioned problem, we can give an overview of existing articles that describe this problem in one or the other way. The study of the problem already goes back to the sixties. In Price and Beaver (1966) it is found that researchers with many

collaborators have high productivity (so, here, collaboration is expressed in terms of number of collaborators and e.g. not in terms of number of co-authored papers). Zuckerman (1967) concludes the predictable fact that Nobel laureates publish and collaborate more than a matched sample of scientists. In Beaver and Rosen (1979) one finds that even in the period 1799-1830, the French scientific elite had a high average productivity in the group of scientists that collaborated. Pao (1992) makes the distinction between global and local collaborators: global collaborators are authors that have co-authors form other laboratoria while local collaborators only have co-authors from their own lab. Pao proves that the global collaborators are much more productive than the local ones. On the other hand, the study Pao (1982) in computational musicology is a bit inconclusive with respect to collaboration and production mainly due to the different sociological habits of the humanities (in comparison with the sciences (see also Pao (1981)). This difference will also be revealed here. The papers Bordons and Gómez (2000) and Subramanyam (1983) comment on some of the above mentioned papers. Also Borgman and Furner (2002) assert that "higher rates of collaboration are useally associated with higher productivity …" without further specifying the definition of "rate of collaboration" and where one even mentions that the assertion may depend on the way productivity is measured. Indeed, especially in collaboration studies it is important to determine whether co-authored papers give a credit of 1 to each co-author (total count) or a credit of 1 divided by the total number of authors of that paper (fractional count) – cf. also Egghe, Rousseau and Van Hooydonk (2000).

From the above it is clear that many different definitions of collaboration and productivity can be given as well as of their relationship. In this paper we define "productivity" of an author as the number of papers written by this author, calculated in the total way. It is our intuitive feeling that calculating authorship in the fractional way cancels the expected higher productivity of authors who collaborate often – although in other contexts, fractional counting is to be preferred, e.g. in the case where one wants to give equal weights (of 1) to each paper (but that does not count in this study).

In this paper, the above defined productivity will be studied in function of the degree of collaboration of this author defined as follows: the fraction of co-authored papers of this author (i.e. the fraction of papers of this author in which this author has at least one co-author). We realize that finer definitions are possible (taking into account the number of co-authors) but our first experiment will be limited to the above definition.

We postulate:

Assertion: In general, the higher number of papers of an author, the higher his/her fraction of co-authored papers.

This assertion was used in Egghe (2008) to refine an earlier model (also given in Egghe (2008)) yielding a law of Lotka for the size-frequency function of co-author pairs, in which the Lotka exponent is higher than or equal to the one of the law of Lotka for individual author size-frequency functions. We will come back to Egghe (2008) in the concluding section after we have studied the two cases in this paper.

A first set of papers under study comprise all papers published in the journal Scientometrics from Volume 1 onwards (and up to 2007 inclusive). The data set has 2,300 papers and 2,003 different authors. We produce the graph: per author: number of papers (counted in the total way) in the abscissa, versus the fraction of co-authored papers in the ordinate.

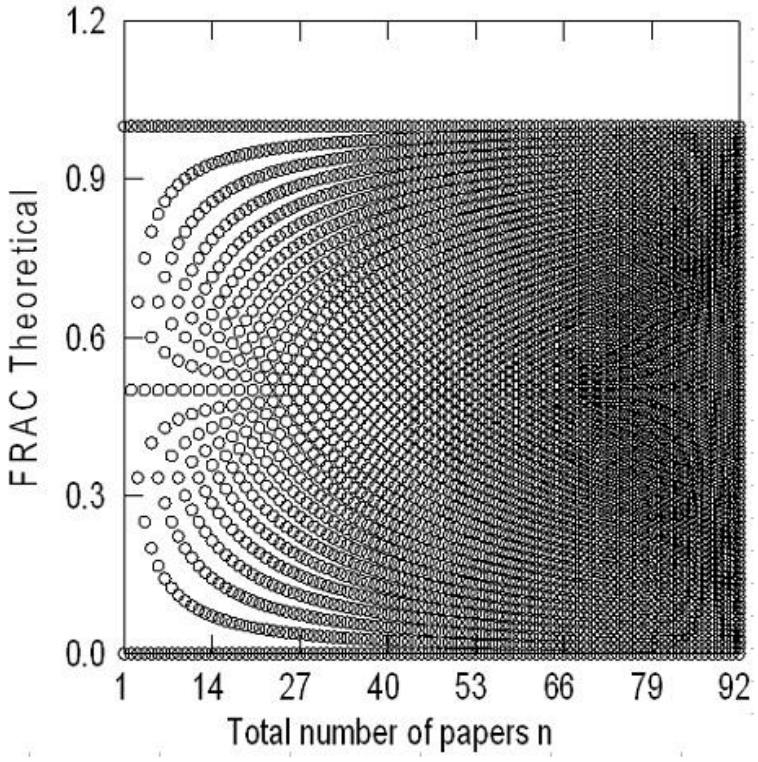

Fig. 1. Theoretical graph of number of publications (abscissa) versus fraction of co-authored papers (ordinate), per author.

Theoretically, such a graph can contain all points of the form $\frac{i}{N}$, $i = 0,1,...,N$ and where N ranges from 1 up to the highest productivity of an author. The theoretical graph of all these trajectories is given in Fig. 1. It is clear that the graph is symmetrical around the horizontal line $y = \frac{1}{2}$.

In the Scientometrics case, obviously, not all points of the theoretical graph are present but we find an almost symmetrical graph around the horizontal line $y = \frac{1}{2}$ based on the data of the 2,000 authors (99.85% of the authors) with less than 60 papers per author.

The remaining 3 authors (66, 84 or 92 papers) show a higher fraction of co-authored papers. These three authors are the Editor-in-Chief, the Editor and the Co-Editor of the journal Scientometrics. It is interesting that a high percentage of their co-authored papers are the results of collaboration among themselves.

We have counted the number of publications of the corresponding co-author pairs:
Number of co-authored papers:

- pair 1: 41
- pair 2: 38
- pair 3: 28

The co-author triple has published 20 papers!

The Scientometrics case shows a normal collaboration pattern in the sense that one has a more or less evenly distributed author pattern of high and low collaboration over all productions and hence the increasing relation, mentioned in the assertion above, is not found. This is executed in the next section where we also present graphs of the production of an author versus his/her fraction of co-authored papers where there are at least 2,3,4,… additional co-authors.

In section III we present the same exercise on the articles in the UHasselt repository. This data set comprises 7,604 papers and 805 different authors.

Now we have that high productivity gives high fractions of co-authored papers but we still have that, amongst the authors with low productivity, we can have low as well as high fractions of co-authored papers, confirming partially the assertion. In any case we have that collaboration as well as productivity is much higher than in the Scientometrics case (the UHasselt repository consists, majorly, of papers in the sciences): most data points are above the horizontal line $y = \frac{1}{2}$. We find an almost symmetrical graph when productivity is matched versus the fraction of papers with at least 4 co-authors in total, a remarkable fact.

The fact that, in the repository case, collaboration and productivity are much higher than in the Scientometrics case is, in itself, also a confirmation of the proposed assertion. The explanation can be given by the different "environments" of both data sets (social sciences (mostly) for Scientometrics and the sciences (mostly) for the repository case).

Further interpretations of these graphs and some open problems are formulated in the last section IV.

II. Collaboration and productivity for the journal Scientometrics

We gathered all papers published in Scientometrics from Volume 1 onwards up to the end of 2007. All authors (2,003) where collected and their number of articles (counted in the total way) were counted (total number of articles was 2,300). We obtained Fig. 2 as the Scientometrics case study of Fig. 1.

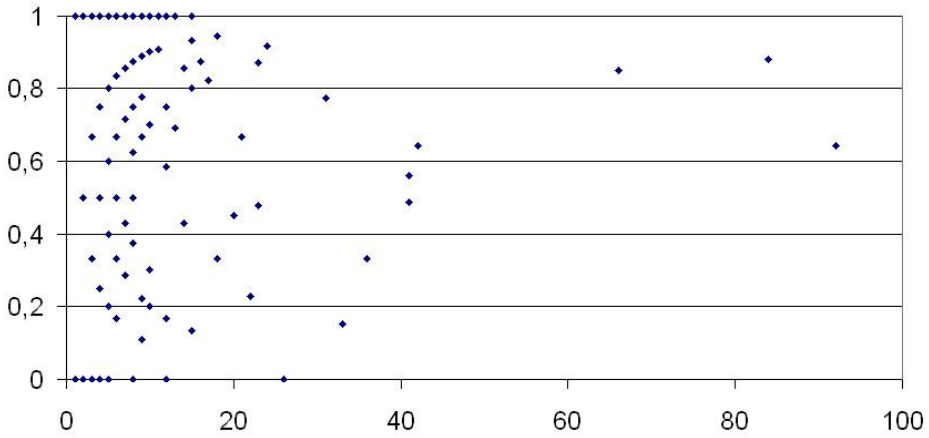

Fig. 2. Scientometrics version of Fig. 1.

We see that the cloud of points is almost symmetrical around the horizontal line $y = \frac{1}{2}$ with exception of the 3 above named editors of the journal. This could be defined as "standard" or "normal" collaboration since there are almost an equal number of high(er) fraction of co-authored papers (above $\frac{1}{2}$) and of low(er) fraction of co-authored papers (below $\frac{1}{2}$). So, within Scientometrics, we do not obtain a confirmation of our assertion.

Figs. 3-5 give the graphs of the same data but now expressing the same productivity versus the fraction of papers with at least 3, 4 or 5 authors, respectively. It is evident that the obtained fractions are decreasing and that they are lower than the ones in Fig. 2. They even start showing a decreasing trend which would indicate (in these higher degrees of collaboration) that the proposed assertion has been contradicted. However we do not think this is true. It is evident that for the higher productions it is not possible to have many papers with a high number of co-authors. For the lower productions and incidental few papers with a high number of co-authors automatically leads to higher fractions.

Figs. 3-5 (and of course also Fig. 2) will become interesting when compared with the repository case which we will study now.

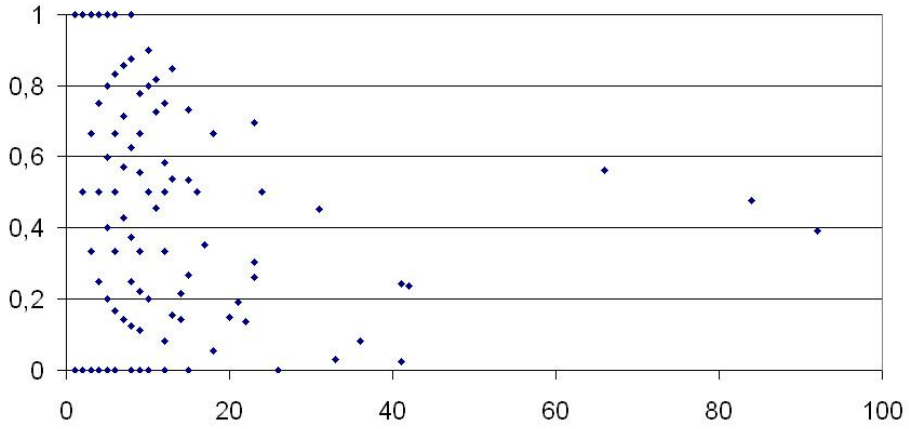

Fig. 3. Scientometrics case of productivity versus fraction of co-authored papers where there are at least 3 co-authors.

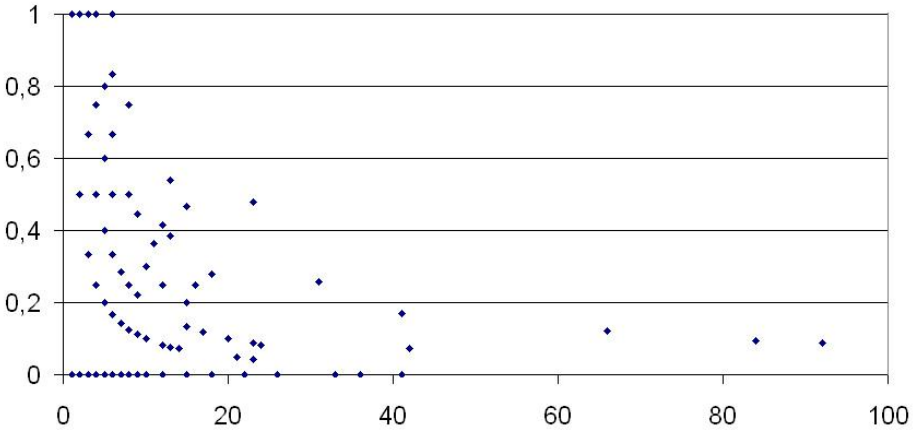

Fig. 4. Scientometrics case of productivity versus fraction of co-authored papers where there are at least 4 co-authors.

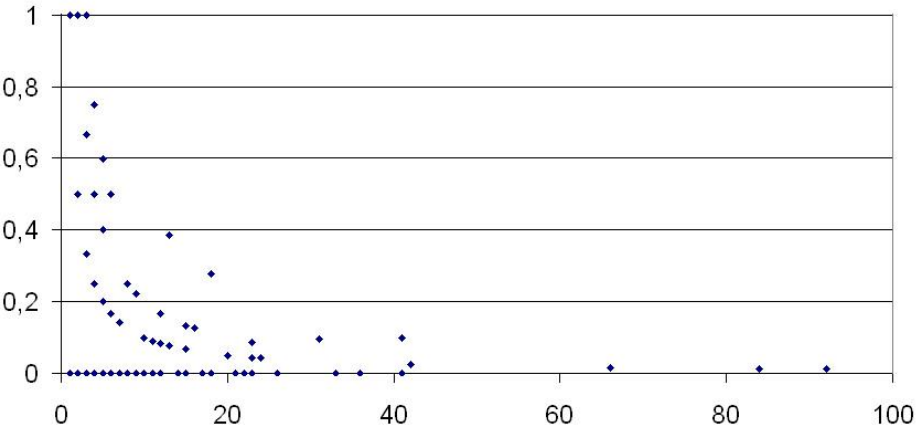

Fig. 5. Scientometrics case of productivity versus fraction of co-authored papers where there are at least 5 co-authors.

III. Collaboration and productivity for the UHasselt repository

Hasselt University is a science and life science oriented university. The department Mathematics-Physics-Informatics has 40% of the papers, Chemistry-Biology-Geology 20%, Medical Sciences 14%, Management, Economics and Law 15% and Social Sciences 11%.

We gathered all papers occurring in the UHasselt (Hasselt University) repository which is continuously updated by the second named author (and by his library collaborators). All authors (805) where collected and their number of articles (counted in the total way) were counted (total number of articles was 7,604). We obtained Fig. 6 as the UHasselt repository case study of Fig. 1.

We immediately notice the higher fractions of co-authored papers at any productivity level. We also notice that the graph is not symmetrical anymore. In fact, most of the points are situated above the horizontal line $y = \frac{1}{2}$, which could be defined as "extensive" collaboration.

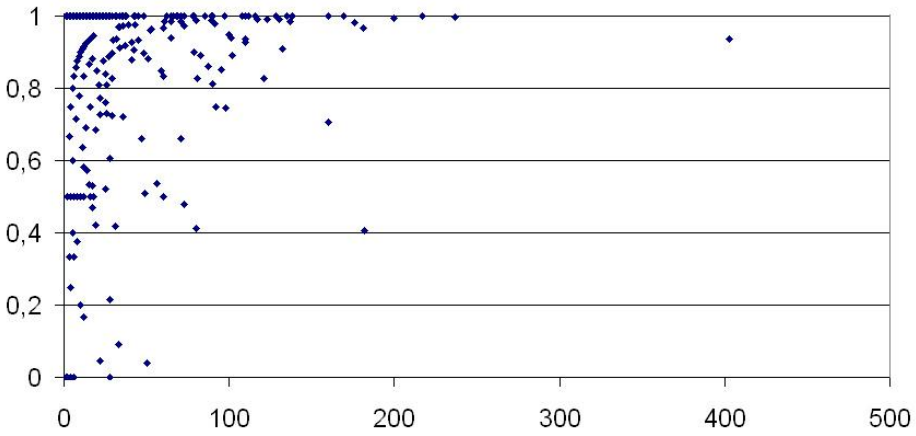

Fig. 6. UHasselt repository version of Fig. 1.

A graph as in Fig. 6 is called a "semi-relation": it is so that, for high(er) production levels we have high(er) fractions of co-authored papers (hence confirming the assertion) but for low(er) production levels we can have low as well as high fractions of collaboration.

In any case, comparison of Fig. 2 and Fig. 6 shows that, in general, production and collaboration are lower in Fig. 2 than in Fig. 6. This finding could be considered as an environmental" explanation and confirmation of the assertion.

The similar graphs as Figs. 3-5 for the UHasselt repository are presented in Figs. 7-9. We see that, although the fractions (evidently) are decreasing from Fig. 7 to Fig. 9 (and also as compared with Fig. 6), collaboration fractions are still very high. It is also remarkable that an (almost) symmetric graph is the one of Fig. 8 where fractions of co-authored papers are considered where there are at least 4 co-authors! This could be defined as "normal" or "standard" collaboration in this case.

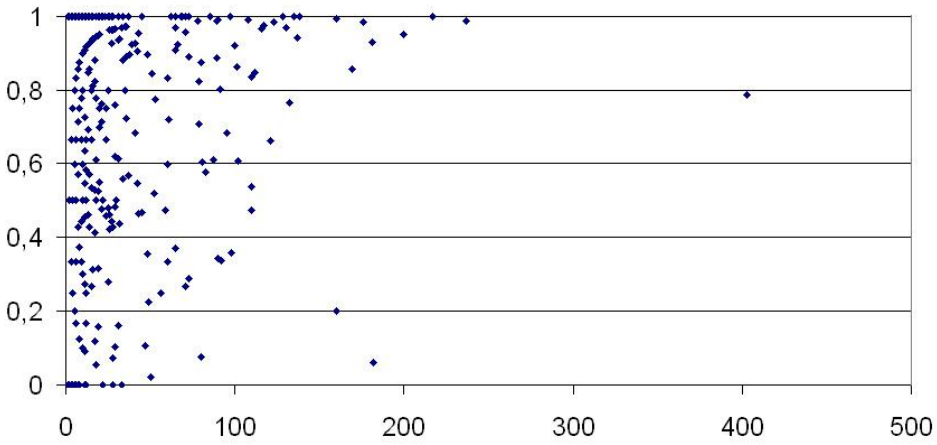

Fig. 7. UHasselt repository case of productivity versus fraction of co-authored papers where there are at least 3 co-authors.

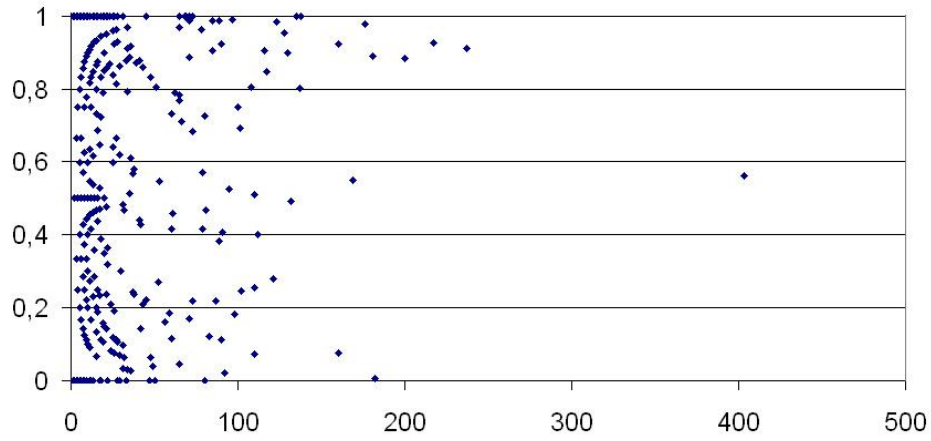

Fig. 8. UHasselt repository case of productivity versus fraction of co-authored papers where there are at least 4 co-authors.

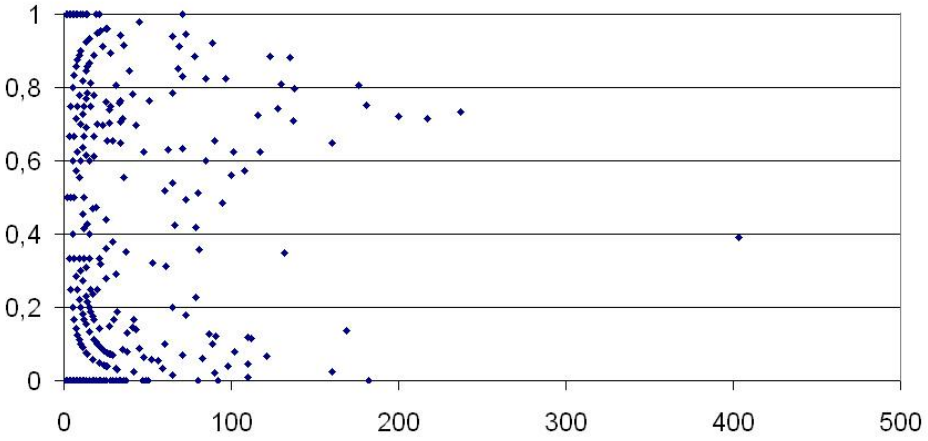

Fig. 9. UHasselt repository case of productivity versus fraction of co-authored papers where there are at least 5 co-authors.

General remark:

All the produced graphs can have points that have a multiplicity higher than one (and hence is not seen on the graphs). This is the more the case where authors have a low production (since there are more authors there). We, however, think that this does not jeopardize our study and or arguments, since multiplicity will occur below as well as above the horizontal line $y = \frac{1}{2}$ in a more or less random way.

IV. Conclusions and open problems

In this paper we studied the assertion that the higher number of papers of an author, the higher his/her fraction of co-authored papers. Two case studies were given: the journal Scientometrics (all volumes) and the UHasselt repository (mainly sciences). The assertion was confirmed by comparing both cases, hence explaining the assertion by the fact that different "environments" have different production numbers per author as well as different levels of collaboration (and where high production goes together with high level of collaboration).

Within the Scientometrics case, however, we did not find a confirmation of the assertion while for the repository case we had a partial confirmation: high production goes together with high levels of collaboration but low production still allows for low or high levels of collaboration.

Based on these results, and going back to Egghe (2008), we can interpret the Scientometrics results as well as the UHasselt repository results in terms of Lotka's exponent for the size-frequency function of co-authored pairs. In Egghe (2008) we showed that, in case the assertion (in Section I) is not valid – more exactly, in case the fraction of co-authored papers of an author is independent of his/her production – that the size-frequency function of co-author pairs is Lotkaian with the same exponent as the law of Lotka valid for individual authors. Hence, in the Scientometrics case, we can conjecture that we are in such a situation, based on the graph in Fig. 2 (fraction of co-authored papers is independent of an author's production). In the case of the UHasselt repository, however, based on the graph in Fig. 6 and the conclusions drawn from it and using the result in Egghe (2008) that, in case the assertion is valid we have a law of Lotka for the size-frequency function of co-author pairs with Lotka exponent larger than the one of the law of Lotka for individual authors, we can conjecture that we are in such a situation here. The confirmation of these conjectures should be given in further experimental research.

We also conjecture that the assertion (Section I) will be less evident in case we count the productivity of an author fractionally, i.e. for every paper of an author we give this author a credit of 1 divided by the number of authors of this paper. Indeed, compared to total counting, the productivity score of an author is smaller and, more importantly, the more co-authors the smaller the (fractional) productivity score, which could make the assertion (section I) invalid. Also this needs further investigation.

References

Beaver D. DeB. and Rosen R. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. Scientometrics 1(2), 133-149.

Bordons M. and Gómez I. (2000). Collaboration networks in science. Festschrift in honor of E. Garfield, ASIST, Chapter 10, 197-213.

Borgman C.L. and Furner J. (2002). Scholarly communication and bibliometrics. In: B. Cronin (ed.). Annual Review of Information Science and Technology 36. Medford, N.J.: Information Today, 3-72.

Borgman C.L. and Furner J. (2008). Personal communication.

Egghe L. (2008). A model for the size-frequency function of co-author pairs. Journal of the American Society for Information Science and Technology, to appear.

Egghe L., Rousseau R. and Van Hooydonck G. (2000). Methods for accrediting publications to authors or countries: consequences for evaluation studies. Journal of the American Society for Information Science 51(2), 145-157.

Pao M.L. (1981). Co-authorship as communication measure. Library Research 2, 327-338.

Pao M.L. (1982). Collaboration in computational musicology. Journal of the American Society for Information Science 33(1), 38-43.

Pao M.L. (1992). Global and local collaborators: a study of scientific collaboration. Information Processing and Management 28(1), 99-109.

Price D.J. De Solla and Beaver D. DeB. (1966). Collaboration in an invisible college. American Psychologist 21, 1011-1018.

Subramanyam K. (1983). Bibliometric studies of research collaboration. Journal of Information Science 6, 33-38.

Zucherman H. (1967). Nobel laureates in science: patterns of productivity, collaboration, and authorship. American Sociological Review 32(3), 391-403.