AUTHOR PRODUCTIVITY AND ERDÖS DISTANCES IN CO-AUTHORSHIP AND IN WEB LINK NETWORKS

Hildrun Kretschmer
[1] NIWI, The Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands
[2] COLLNET, Borgsdorfer Str. 5, 16540 Hohen Neuendorf, Germany
kretschmer.h@t-online.de

Abstract

The increasing scientific-political importance of cooperation in science requires the application of new methods of analysis of social networks in co-authorship and in Web link networks. In this context, some interesting papers on "Erdös Number" which gives the shortest way (topological distance) between an author and the well-known Hungarian mathematician Erdös in co-authorship network have been published recently. A few new queries which particularly concern the position of highly productive authors in the network were developed in the present paper. Thus, a relationship of distribution of these authors among the clusters in the co-authorship network could be proved in dependence on the size of these clusters. Highly productive authors have on an average low Erdoes Numbers and thus shorter distances to all the other authors of a special field than low productive authors whereby the influencing possibility of highly productive scientists gets expressed among others in the development of this special field.
A theory on the stratification in science with respect to the over random similarity of the scientists who are collaborating with one another which could be covered with other empirical methods before could also be confirmed on application of the Erdös Numbers.
The application of the new developed queries also on the web links between homepages of authors is proposed for studies in future. It has to be studied if co-authorship and web link networks have similar structures or not regarding the author productivity and Erdös Distances.

1 Introduction

The increasing cooperation in science, not only in basic research but also in the applied research and technology, is a well-known phenomenon since years. The increasing scientific-political importance of cooperation in science requires the application of new methods of analysis of social networks in co-authorship and in web link networks.
A common bibliometric method for measuring the cooperation is the analysis of co-authorship networks. A suitable webometric method has to be developed in the future.
There are various references to the positive effect of "multi-authored papers" in the co-authorship network: for example several studies show that international cooperation is linked with a higher `citation impact' (Glänzel 2002).

The investigation of these processes can be made by analyses at the micro level (individuals), at the meso level (institutions) or at the macro level (countries) (Glänzel

Hildrun Kretschmer, *Author productivity and Erdös distances in co-authorship and in web link networks*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 393-400.

2002). In the field of scientometrics and informetrics one most frequently comes across studies on international cooperation in science, followed by cooperation relationships between institutions.

On the other hand, bibliometric analysis which have networks of individuals to objects are essentially very rare to find (Newman 2001). But because the knowledge at meso and macro level does not yet give adequate statements on the `trends' in the cooperation between individuals, it is necessary to carry out investigations at the micro level to an increasing extent in the future.

2 Co-authorship Networks at Micro Level and Erdös Numbers

In this context it deals among others with the question as to how close are scientists of a scientific field connected with one another in the co-authorship networks. The closer the connection and the greater the network the higher will be the speed and range of information transfers within the scientific community, i.e. the scientific results are no longer delivered by the individual scientists but by the total network as a whole (Newman 2001).

This can be understood in the following manner: Under the assumption that the exchange of information between two co-authors A and B is particularly extensive and deep because of personal contacts one can further presume that a part of this information also reaches C if B is in co-authorship with C, also if C is not the co-author of A. The same also holds good for the information flow in the direction D in the case of a co-authorship between C and D. This principle can be further continued in the same manner. The information disseminates via such chains of co-authorship in both the directions whereby a mutual scientific influencing of the authors is also linked.

Some interesting works/papers for analysis of co-authorship networks in the special field of science, which are important to be further worked out, have been published recently. An example is the application of Erdös number (Genest & Thibault 2001, Balaban & Klein 2002). Erdös was a very famous Hungarian mathematician who excelled among others through his extensive cooperation with a large number of other scientists, whereby he exercised a great influence on the development of research in his field.

The computation of Erdös number is very simple: Erdös himself obtained the Erdös number EN=0. All the authors who have published at least one paper together with Erdös obtain the number EN=1. All the authors obtained EN=2 who have not published along with Erdös, but at least with one of his co-authors (EN=1). This principle is continued, i.e. all the co-authors who have not published with Erdös and with none of his co-authors together but with at least with one of the authors with the Erdös number EN=2, obtain the Erdös number EN=3, etc. In this way, for example, an author with EN=10 is linked with Erdös through a chain of co-authorships.

Therefore, each scientist in the field of mathematics can ask for his own Erdös number EN which indicates the shortest path (topological distance) between him and Erdös. Two samples of scientists are resulting.

First sample (including Erdös): EN=d can be determined for each scientist of this sample. Following chains of co-authorships exist between all the authors of this first sample. Therefore, these authors form a cluster.

Second sample: But no Erdös number EN can be determined for a scientist who was not in co-authorship with any of the authors of the first sample, i.e. he belongs to a different cluster.
Therefore, among others investigation was done so far to find out how many clusters are there in a data set and what scope these clusters have (e.g. in a special field of science)

3 Derivation of New Queries

Till now rather only known and highly productive scientists of a data set which is to be analysed were mainly in the center of investigations, i.e. the value EN=0 was assigned to them, similar like to Erdös.
Therefore Genest and Thibault (2001) have recently proposed to extend this method in future investigations and to determine the Erdös distances (here: ED) in large clusters, i.e., the shortest path (topological distance) between two randomly selected scientists, and to compute the average there from.
This average Erdös distance is analogous to one of the measures of centrality which are known from the Social Network Analysis (SNA) and indeed the "Closeness Centrality". Otte and Rousseau (2002) have already referred to the increasing significance of SNA for investigations in the information science. Regarding SNA, see also Wasserman and Faust (1994), Batagelj, V., Ferligoj, A., and Doreian, P. (1992).

By additional incorporation of the observation of Braun, Glänzel and Schubert, (2001) that so far there are only a few bibliometric investigations on the relation between the productivity of the authors and their cooperation, the idea originated in the case of the author of the present paper to check whether there exists a relationship between the average of Erdös distances and the productivity of the authors.

Based on the sociological perspective, Genest and Thibault also proposed in 2001 it would be useful to investigate whether the Erdös distances between the publishing researchers can be related to an earlier published concept of Kretschmer (1997). Genest and Thibault suggest if this should be the case then this can be evaluated as confirmation of Kretschmer´s concept.
According to this concept a special kind of social stratification in science can be proved. Here social stratification means that the probability of *personal* contacts like friendship or cooperation between similar scientists is higher than between dissimilar (different) ones (Birds of a feather flock together) and also that this probability decreased with an increase in the dissimilarity between the scientists. Similarity relates to various personality characteristics, for example to the age or here to the productivity.
There after it would be expected in the present paper that this kind of social stratification can be proved related to productivity and low Erdös distances. But this stratification becomes weaker and weaker with the increase in the ED because of the decreasing frequency of *personal* (face-to-face) contacts.
Regarding the proposed webometric study in future such stratification pattern can be expected under the condition the probability of links between homepages of scientists is increasing with increasing frequency of *personal* contacts. This assumption is consistent with findings that web links represent relatively informal scholarly communication (Wilkinson, Harries, Thelwall & Price, 2003). According to this scheme the reason for links between homepages of a member of a research team and the homepage of one of the other members is based in research partnership.

Therefore in conclusion the following assumptions are checked:

1. There is a connection between the structure of clusters and productivity of scientists.
2. There exists a relationship between the average of Erdös distances and the productivity of scientists.
3. A social stratification can be proved in relation to low Erdös distances whereby the visibility of this stratification decreases with increasing Erdös distances.

4 Data

The new special field of physics which was established by K. von Klitzing in the year 1980 with the discovery of Quantum-Hall effect is investigated in the study in the period 1980-1985. The publication data were determined by H.J. Czerwon (1993) and analyzed by him under a different aspect than in the work presented here.
381 documents including 385 authors were identified by Czerwon in the INSPEC data bank on Quantum-Hall effect for the period from 1980-85 (Full details are available after request).

5 Methods

The Erdös distances (ED) were determined between all the possible pairs of the 385 authors of the whole data set. The Pajek program was employed for this purpose.
The clusters and their sizes were determined on the base of these existing Erdös distances.

In a second step the authors were grouped according to their productivity, i.e., corresponding to the number of their publications i per author (Normal count procedure: Each time the name of an author appears the name is counted).

In order to avoid statistical fluctuations, the data are classified according to the logarithm of the number of papers (class X=1 contains those authors with i= 1 publication per author, class X = 2, authors with 2-3 publications, X = 3, authors with 4-7 publications, X = 4, authors with 8 and more publications).

The distribution of the Erdös distances and the average was separately determined for each class of authors and then the distributions and the averages of the different classes were compared with one another.

There are various methods to identify the stratification. A special interaction index H_{XY}, which is well known in sociology (Wolf 1996, called here: homophilie index) was employed in the case of Erdös distances. Thus here it deals with the ratio of the observed to the statistically expected frequency of occurrence of ED_{XY} between the class of authors with X publications per author and the class of authors with Y publications per author:
The special interaction index H_{XY} is defined as:
$H_{XY}=ED_{XY}/(G_X \bullet G_Y/G)$
where G- geometric mean of all matrix data ED_{XY}
G_X – geometric mean of the data in row X
G_Y – geometric mean of the data in column Y
In case of stratification H_{XY} must be higher between the authors with same or similar productivity (X=Y) than between authors with different productivity ($X \neq Y$). The special interaction index H_{XY} was separately employed at various Erdös distances, i.e.

on the one hand at the distribution of ED=1, on the other hand at distribution of ED=2, etc. The results were compared with each other.

6 Results and Discussion

1. The largest cluster covers almost 40% of the 385 authors. In addition there are still a large number of small and very small clusters, see Table 1. This structure of clusters which contain a single very large cluster and also a large number of small clusters, is in agreement with the existing publications in the literature (Newman 2001, Genest &Thibault 2001). It is possible this could denote a general rule in any co-authorship network.

Table 1: Number of Clusters

	Number of authors within one and the same cluster:									
	1	2	3	4	5	6	8	10	13….	**144**
Number of Clusters	49	21	16	10	1	2	1	2	1….	**1**

The highly productive authors (X=4) thus find themselves mainly in the very large cluster whereas the low productive authors (X=1) can be encountered relatively frequently in the middle and very small clusters (Fig.1). The left diagram in Fig.1 shows the percentage distribution of the highly productive authors (X=4) among the clusters. 76% of the highly productive authors are in the largest cluster (inner slice). Only 24% of the authors belong to the medium and smaller clusters (outer slice). The right diagram shows the percentage distribution of the low productive authors (X=1). While only 28% of these authors can be found in the largest cluster (inner slice), 72% belong to the medium and smaller clusters (outer slice).

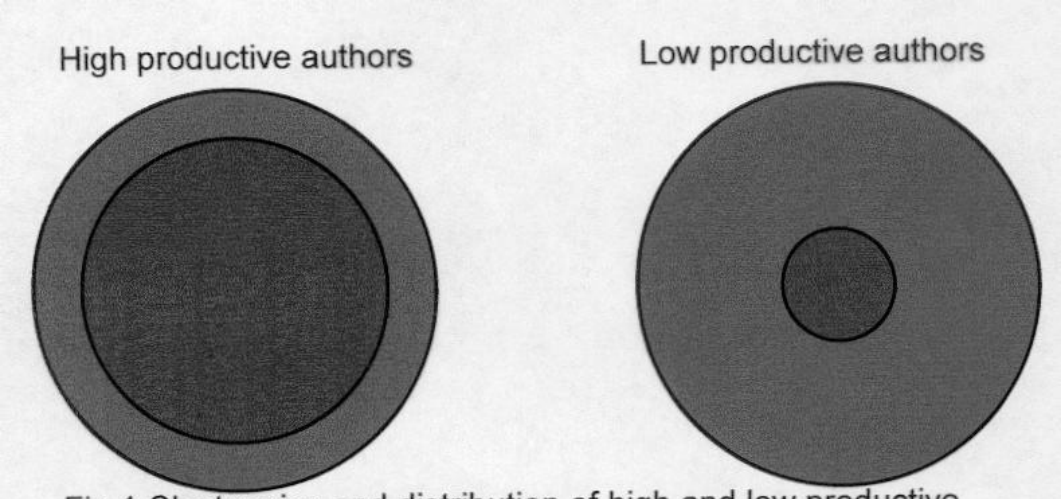

Fig.1.Cluster size and distribution of high and low productive authors. Percentage of authors in the largest cluster, see inner slice

Thus, the complete structure of clusters indicates the high degree of collectivity in the scientific work as well as the special role of high-productive authors.

2. All the authors of the complete data set, i.e., independent of their affiliation to different clusters are considered in the following (Fig. 2). Analogous to Erdös, each author was assigned the Erdös Distance ED=O and the number of authors was determined which can be assigned to him with the Erdös Distance ED=1, subsequently the number of authors with ED=2, etc. A distribution of the number of authors corresponding to the Erdös Distances ED=d resulted there - from for each author with ED=O. In each class of authors, the respective average per author was computed for each Erdös Distance ED=d. These average distributions of various classes of authors were compared with each other (Fig 2). The Erdös Distances ED=d are present in the abscissa and the respective average of the relevant number of authors in the ordinate. While in the case of highly productive authors (X=4) an Erdös Distance of ED=3 can be assigned as median (Fig. 2, on the left), in case of low productive authors (X=1) an Erdös Distance of ED=4 can be assigned as median (Fig. 2, on the right).

In general the averages of Erdös distances are different between the classes of authors having varying productivity X. Authors with higher productivity have on an average lower Erdös distances to all the authors of the data set than the groups of authors with lower productivity.

This indicates that high-productive scientists have a greater influence on the entire scientific community than the low-productive authors. This phenomenon is in accordance with the earlier studies related to other bibliometric measures than co-authorships. For example, Bakker and Rigter (1985) could prove a high correlation between the scientific productivity, citation rate and editorship of influential journals.

High-productive authors also have in average relationships with a higher number of authors on the whole in the form of Erdös distances than the low-productive authors (Fig. 2). This is determined by the higher affiliation of the low-productive authors with the small and very small clusters.

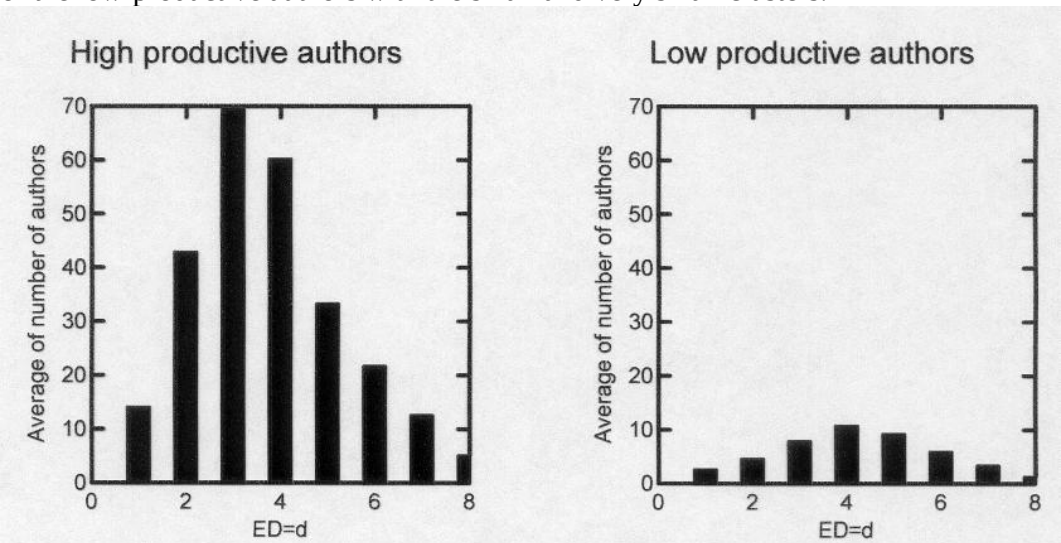

Fig.2. Distributions of averages of number of authors

3. Considering only the largest cluster, separated from the other clusters, then the authors with lower productivity get differentiated by the fact that the authors who have a small Erdös distance to a high productive author, also have smaller Erdös distances to

all the other authors on the whole than those authors with lower productivity who have a high Erdös distance to a highly productive author (Fig. 3). Therefore, a query regarding the role of high-productive scientists should be made at this point in the future investigations for the promotions of young scientists by high productive authors.

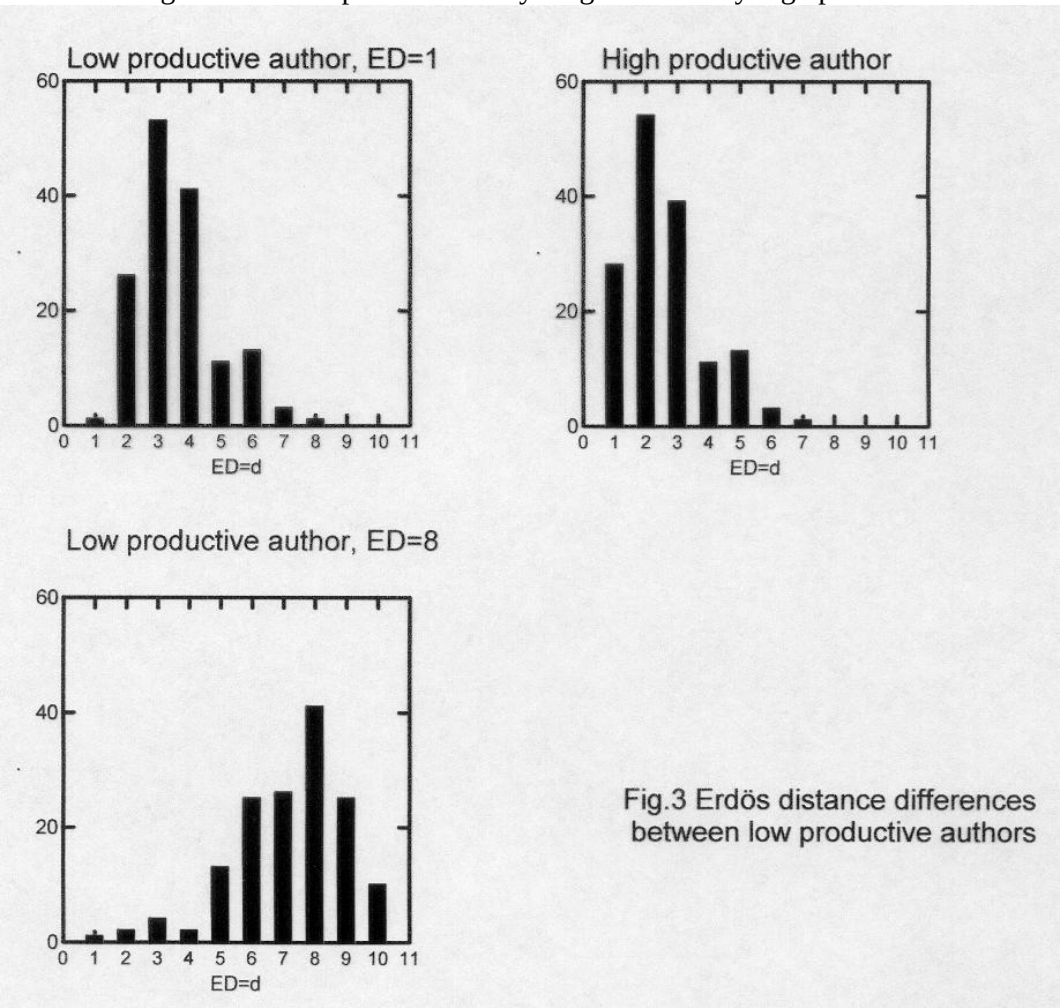

Fig.3 Erdös distance differences between low productive authors

The right upper diagram in Fig. 3 shows the distribution of authors according to Erdös Distances ED=d related to a highly productive author (X=4) with ED=O. The two other diagrams in Fig. 3 have to do with two low productive authors (X=1) with ED=O. Whereby one of them (diagram, above left) has a low Erdös Distance ED=1 related to the highly productive author, and the other (diagram below it) has a high Erdös Distance ED=8. In case of the latter the median also accepts a higher value than in the diagram above it.

4. Authors who have published a common paper (ED=1) usually know each other (Except co-authors of articles with more than about 50 co-authors). Certain personal relationships are also formed frequently with the co-author of a co-author (ED=2), but mostly less pronounced than with the co-author himself. This weakening increases with the increase in Erdös distances between the authors. Therefore, the above named special kind of social stratification of the authors can be proved in the distribution of the frequencies of occurrence of the Erdös distances ED=1 between the classes of authors and diminished also in the distribution of the frequencies in occurrence of the

Erdös distances (ED=2). As expected this effect is no longer present in the higher Erdös distances whereby Kretschmer´s concept is confirmed (Fig. 4).

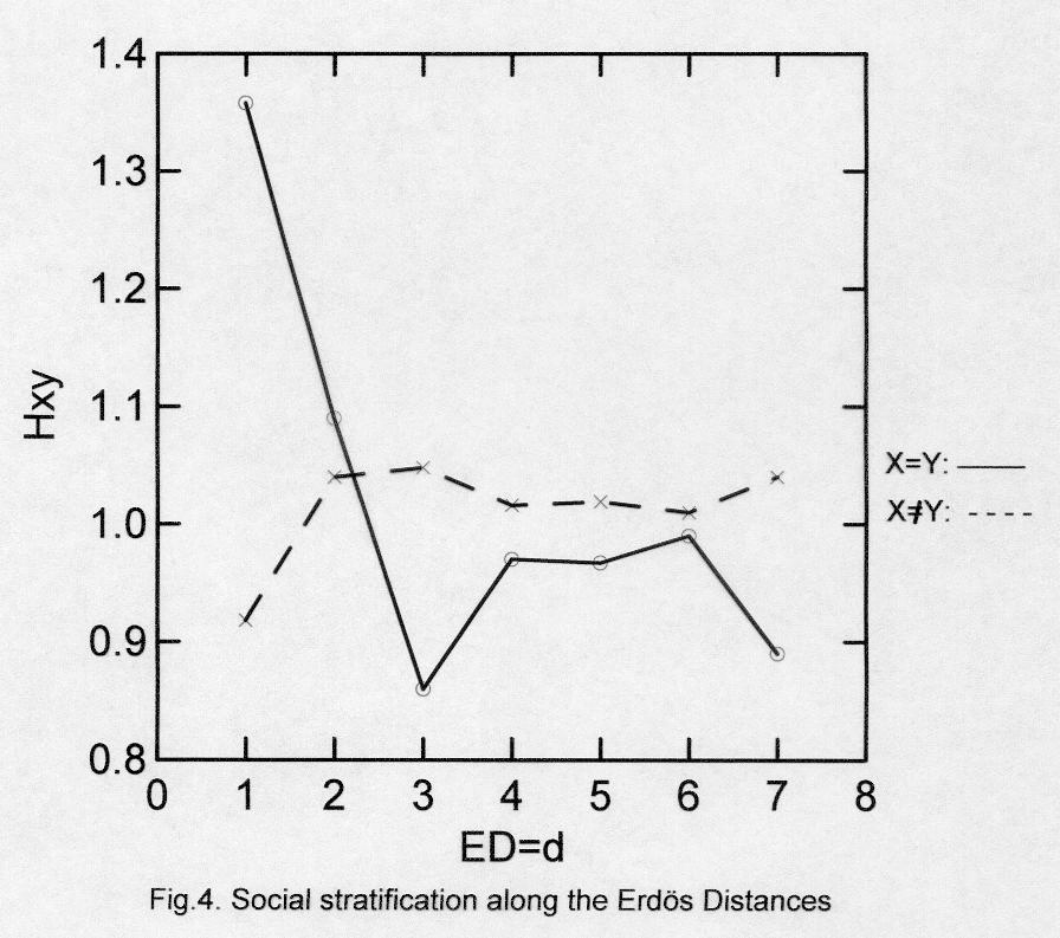

Fig.4. Social stratification along the Erdös Distances

The special interaction index H_{XY} was separately employed at various Erdös distances, i.e. on the one hand at the distribution of ED=1, on the other hand at distribution of ED=2, etc.
Like announced above in case of stratification H_{XY} must be higher between the authors with same or similar productivity (X=Y) than between authors with different productivity (X≠Y).
In Fig. 4 the Erdös Distance ED is present in the abscissa, and the interaction index H_{XY} in the ordinate. The curve with the full line has to do with the values for the interaction index H_{XY} between the classes of authors with the same or similar productivity (X=Y) and the curve with dotted line has to do with the values for interaction index between classes of authors with different productivity (X≠Y).
As presumed, the interaction index in ED=1 is distinctly higher between authors with same or similar productivity than between authors of different productivity. This trend is in reduced form in ED=2, but no longer in case of the higher Erdös Distances.

7 Proposal for Further Investigations

Further investigations on the changes in the clusters are promising, particularly the query as to what extent a sudden growth in one of the small clusters could be connected with the development of a new special field.

Like mentioned above the increasing scientific-political importance of cooperation in science requires the application of new methods of analysis of social networks not only in co-authorship but also in web link networks.
The EU has recently financed a new consortium from England, The Netherlands and Spain to investigate further the potential to create new indicators from the web for use in science and technology policy making. This is a three-year project that started in November 2002 and is one possible direction for the future of information and communication science research. The WISER project will provide useful resources for those wishing to start webometrics and information on the potential of a range of new webometric techniques (Kretschmer & Thelwall). There is a proposal in this line for the application of the new developed queries in the present paper also on the web links among homepages of authors. In a first pilot study both the co-authorship network and the web link network among 2000 members of the German Society of Psychology have to be studied. There is the question if both networks have similar structures or not related to author productivity and Erdös Distances and which conclusions can be drawn for science policy.

8 Acknowledgements

This work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015) (www.webindicators.org).

References

Bakker, P. & Rigter, H. (1985). Editors of Medical Journals: Who and from where. Scientometrics, 7, 11-22

Balaban, A. T. & Klein, D. J. (2002). Co-authorship, rational Erdös numbers, and resistance distances in graphs, Scientometrics, 55, 59-70

Batagelj, V., Ferligoj, A., and Doreian, P. (1992). Direct and indirect methods for structural equivalence, Social Networks, 14, 63-90

Braun, T., Glänzel, W. & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. Scientometrics, 51, 499-510

Czerwon, H-J. (1993). Bibliometrische Verteilungen und die Evolution von Frontgebieten der Grundlagenforschung. In: W. Neubauer & K.-H. Meier (Eds.), Deutscher Dokumentartag 1992 (pp. 625-640). Frankfurt am Main: Deutsche Gesellschaft für Dokumentation e.V.

Glänzel, W. (2002).Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. Library Trends, 50, 461-473

Genest, C. & Thibault, C. (2001). Investigating the concentration within a research community using joint publications and co-authorship via intermediaries. Scientometrics, 51, 429-440

Kretschmer, H. (1997). Patterns of behaviour in co-authorship networks of invisible colleges. Scientometrics, 40, 579-591

Kretschmer, H. & Thelwall, M. (2003). The Development of Information Professionals: The European Perspective - The Way from Librametry to Webometrics. MALA: Madras 2003, 13-25

Newman, M. (2001). The structure of scientific collaboration networks. Proc. Natl. Sci. USA, 98, 404-409

Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science, 28, 443-455

Wasserman, S. & Faust, K. (1994). Social network analysis. Methods and applications. Cambridge: Cambridge University Press 1994

Wilkinson, D., Harries, G., Thelwall, M. & Price, L. (2003). Motivation for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication. Journal of Information Science, 29, 59-66

Wolf, Ch. (1996). Gleich und gleich gesellt sich. Individuelle und strukturelle Einflüsse auf die Entstehung von Freundschaften. Hamburg: Verlag Dr. Kovac

A revised version of this contribution has been published as:

Hildrun Kretschmer (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. Scientometrics, 60(3), 409-420.