A Methodological Study for Measuring the Diversity of Science

MARION SCHMIDT[1]
JOCHEN GLÄSER[2]
FRANK HAVEMANN[1]
MICHAEL HEINZ[1]

1 Department of Library and Information Science
Humboldt University Berlin, D-10099 Berlin, Germany
2 Research Evaluation and Policy Project, Research School of Social Sciences
The Australian National University, Canberra ACT 0200, Australia

Abstract

Inspired by the hypothesis that diversity of research might decline as a result of new science policy measures (e.g., evaluation-based funding), we seek to explore bibliometric methods to analyse the structure of research landscapes. The task is to make quantitative changes in the diversity of research landscapes measurable and therefore comparable in time series as well as between different countries.

1. Introduction

We are currently observing an increasing interest in the diversity of science, which is considered to be under threat by science policy's search for excellence. The dominant strategy for achieving excellence is to concentrate funding on the best performers. This strategy implies that fewer units will be funded in each field, which is likely to diminish diversity (Adams [1]).[1] At a more subtle level, diversity is threatened by the adaptive behavior of scientists. Whenever science policy increases punishment for failure, e.g. via reduced funding, researchers are likely to choose projects that are safe in that they are approved of by the scientific community and have a high probability of success. Such safe projects follow the mainstream of a field and use approaches that are known to yield results.

These arguments, albeit persuasive, lack empirical foundation. No convincing measurement of research diversity has so far been provided. Empirical studies on the impact of the British Research Assessment Exercise (RAE) on the diversity of economic research report that in the opinions of researchers there is a homogenisation (Harley [2]). These opinions, while suggestive, cannot be regarded as reliable evidence for two reasons. Firstly, perceptions of a changing diversity depend on scientists' individual scientific perspectives and their opinions about science policy. They may therefore be biased. Secondly, quality and marginality of a scientific enterprise are often inseparable. Nonconformist approaches might be perceived as bad science by the majority. Conversely, scientists might rationalise insufficient recognition of their work as being due to the specificity rather than quality of their work.

[1] Cf. also J. Molas-Gallart, A. Salter, Diversity and excellence: considerations on research policy, IPTS Report, 2002, Vol: 66. Available online: http://www.jrc.es/pages/iptsreport/vol66/english/ITP1E666.html

In order to test the 'homogenisation thesis' described above, we need measures of diversity that do not depend on the perception of scientists. Bibliometric indicators can be used to construct these measures because they are unobtrusive and objective, i.e. they neither affect the behaviour they measure nor depend on scientists' opinions about the attribute that is measured. To our knowledge, the first one to propose a bibliometric approach to measuring diversity was Hariolf Grupp [3]. We follow the proposal by Jonathan Adams (University of Leeds) at the 8th Science and Technology Indicators Conference in Leiden 2004 to use the concept of diversity developed in ecological research as a starting point for the analysis of the diversity of science. However, because of the inseparability of marginality and quality we are reluctant to use impact indicators in the measurement of diversity, as he has proposed.

This paper reports the results of a feasibility study that uses measures of diversity commonly used in biodiversity research and applies standard techniques (co-citation cluster analysis) to a field that shows relatively clear natural boundaries, namely Electrochemistry. The aim of this study is to establish if co-citation clusters and the publications citing these clusters can be identified with reasonable efficiency using ISI's Science Citation Index on CD-Rom, and how time horizons must be defined in an analysis of the dynamics of diversity. In addition to the results presented we briefly discuss other methods to detect structures of research fields.

2. *Approach*

2.1. *The concept of diversity*

The concept of 'diversity' has rarely been used in science studies and has not yet been defined with sufficient rigour. In order to investigate the political concerns described in the introduction, we apply the concept to scientific fields, which we regard as consisting of several approaches. Approaches may be concerned with the application of a specific method, the investigation of a specific object, the application of a specific theory, or any combination of the previous. They may be complementary (indifferent or symbiotic), or they may contradict each other.

We would intuitively consider a scientific field that consists of more approaches as being more diverse. The political concerns address precisely this question. Does the number of different approaches in scientific fields reduce because scientists respond to science policy interventions by switching to the mainstream?

In order to proceed from the intuitive understanding of diversity to a more definitive and measurable understanding we can draw on the research on biodiversity. In this research, the concept of diversity is usually linked to "the variety and abundance of species in a defined unit of study" (Magurran [4], see also Gaston [5]). From this follows that biodiversity is characterised by two basic measures. The oldest and most intuitive measure of biological diversity is simply the number of species in the unit of study or "species richness" (Magurran [4]). The 'evenness' of a unit describes the variability of species abundance. These measures can be combined in a 'diversity index', i.e. in "a single statistic that incorporates information on richness and evenness" (Magurran [4]). Among the various diversity measures, the Shannon index is one of the most enduring (Magurran [4]).

For the purpose of this study, we identify the approaches in a scientific field as 'research fronts', i.e. as a group of papers that refers to the same co-citation cluster. These research fronts can be regarded as 'species' in the units we analyse (the fields at the international and national levels). Research articles belonging to a research front can be regarded as the individuals belonging to a species.

Applying the ideas of biodiversity research, we will consider two measures of diversity, namely 'species richness', i.e. the number of research fronts, as the simplest measure of diversity, and the Shannon index (entropy) as a synthetic measure of 'research front richness' and 'evenness'.

Since we want to explore the suitability of bibliometric indicators for measuring the diversity of a country's research base as it develops over time, and since a country's research base is an inseparable part of international science, relative measures are needed that relate the national diversity to the diversity of the international scientific field. Assessing the diversity of a research field thus requires the delineation of fields at both the international and the national level. The field level is important because diversity is assumed to affect the production of knowledge, which takes place in smaller collectivities (international scientific specialties). However, field delineation is one of the most difficult tasks, and is currently regarded as unsolved in bibliometrics (van Raan [6]). The major methodological problem is that due to the overlap of fields and the skewed distribution of participation by scientists, bibliometric indicators don't produce visible boundaries (Noyons & van Raan [7]).

Therefore, new methods for delineating fields need to be explored. However, the delineation of fields does not produce a major obstacle in this project. Since we are looking for a relative measure of diversity, any error in field delineation that affects the international level and the national level in the same way will not distort the measurement of diversity. Adverse effects can be expected only in the extent to which part of a country's research is not at all published in journals that can be identified as belonging to a field at the international level. Since our study was focused on the feasibility of diversity measures, we selected a field that shows relatively clear natural boundaries, namely Electrochemistry. By choosing such a field, the issue of delineation can be circumvented, and the suitability of measures of diversity can be explored.

A second task that follows from our approach is the identification of 'species', i.e. research fronts, which need to be applied at both international and national levels. A first candidate for this measurement is co-citation analysis, with highly cited and highly co-cited papers being defined as a set of scientific works to which research orients over a longer period of time and which therefore can be assumed to constitute a distinct approach or perspective. The number of distinguishable co-citation clusters can be assumed to be an indicator of diversity at the level of international fields, and the number of those clusters to which a national sub-field contributes can be assumed to be a measure of national diversity. Both measures have to be constructed in such a way that diversity can be measured in a short period of time, thus enabling a dynamic analysis of the development of diversity over time.

2.2. Database

We built a database with all 4522 records (articles, reviews etc.) from 14 journals in Electrochemistry from the SCI 1998. The Electrochemistry journals were selected from three sources: (1) Leydesdorff's cluster analysis of SCI journals,[2] (2) a search in Web of Science (WoS) with string `electroch* or elektroch*`,[3] and (3) the ISI journal list.[4] The three lists were then matched with the SCI 1998. The records were drawn from CD-ROM edition of SCI 1998. The journals selected and their record numbers can be seen in Table 1. We found 4257 articles, 62 letters, and no notes in this dataset, all together there were 4319 so-called research papers, of which 110 were without references.

2.3. Methods

We first performed a co-citation cluster analysis by means of a single linkage cluster routine, which is scalable for large datasets and can be applied in combination with any proximity measure. We constituted so-called research fronts by projecting the co-citation clusters to the current level. After that we measured the distribution of research fronts by the Shannon entropy formula and calculated the contributions of six countries to this distribution and then measured the diversity of these countries' research landscapes.

[2] http://users.fmg.uva.nl/lleydesdorff/jcr01/c55.htm, retrieved 2004-12-9

[3] http://isi1.isiknowledge.com, retrieved 2004-12-9

[4] http://sunweb.isinet.com/cgi-bin/jrnlst/jlresults.cgi? PC=K&SC=HQ, retrieved 2004-12-9

Table 1: Numbers of records in SCI 1998 of 14 journals in Electrochemistry and numbers of their research papers (SCI document types *article, letter, note*)

JOURNAL	RECORDS	RESEARCH PAPERS
Bioelectrochemistry and Bioenergetics	129	122
Chemical Vapor Deposition	51	37
Corrosion Science	151	146
Electroanalysis	245	236
Electroanalytical Chemistry	4	0
Electrochimica Acta	551	536
Journal of Applied Electrochemistry	147	144
Journal of Electroanalytical Chemistry	706	693
Journal of Power Sources	446	437
Journal of the Electrochemical Society	735	711
Plating and Surface Finishing	217	136
Russian Journal of Electrochemistry	239	228
Sensors and Actuators B - Chemical	333	331
Solid State Ionics	568	562
Sum	4522	4319

We chose the well-known *Salton's Cosine* as similarity measure to normalise the raw co-citation counts in order to balance highly cited papers and less highly cited papers. As purported in the recent debate in *JASIST*, the other standard measure for interval scaled data, the often used *Pearson's r*, implies certain theoretical problems. Since it measures the amount to which a linear function exists between two variables, it should only be applied to datasets that are normally distributed, whereas bibliometric data are skewed. Apart from this formal objection, it can be argued that it is more adequate to measure similarity by normalising the exact co-citation count than measuring the existence of a linear relationship.

Leydesdorff & Bensman [8] showed that the logarithmic transformation, which is often suggested to make the data conform to the requirement of normality, reduces the variance and therefore must be considered as inappropriate for any classificatory purpose. Besides, Ahlgren et al. [9] showed that *Pearson's r* behaves in an inadequate manner in the case of adding zeros – which means a situation in which another dataset is added to an existing dataset with no correlations between them.

We started with a combination of an integer citation threshold, a fractional citation threshold and the cosine-normalised co-citation threshold. However, it was impossible to achieve an adequate coverage using three thresholds. Former studies like the one of van Raan [10] and the studies performed by Small & Sweeney [11] that use such a combination of thresholds only cluster a small amount of highly cited and co-cited papers, whereas we considered it to be problematic to statistically evaluate such small data distributions especially when applied to a field level.

So we affixed the citation count at $c > 1$ and focused on the co-citation threshold, that was gradually increased. Peaks in the increase of the number of clusters and in the parallel decrease of the coverage were used to fix a cluster distribution. The co-citation clusters were projected onto the current level of the *SCI* 1998 edition. We chose to set the very simple condition that articles have to cite at least one reference from a co-citation cluster. Every article that is part of a research front is bibliographically coupled with at least one other article in that research front.

Additionally, we contribute to a high coverage with our decision to include research fronts of those references that have been cited at least two times but haven't been clustered at the chosen co-citation threshold. This was done in order to reflect more adequately the current research landscape – we wanted to complement those research fronts whose development reaches back in time by those which evolve in the present and therefore can't be reflected in co-citation clusters. So, essentially we use a combination of co-citation and bibliographic coupling.

Our approach allows overlapping research fronts. As Jarneving [12] mentions, one can regard the research fronts that are not disjoint as bibliographically coupled on a higher level, but in order to measure the diversity of the field it is not possible to incorporate group memberships on different levels. If we simply counted articles several times when they are part of several research fronts (as Jarneving does), the proportion of articles in research fronts to the total amount and to the ones that are not in research fronts would be biased. So, we decided to assign these overlapping articles fractionally to the research fronts. This means, that if an article cites two clusters, the value of 1/2 would be assigned to both research fronts, respectively.

To measure the diversity of the research landscape of electrochemistry on a global level the number of research fronts was counted, and the entropy formula was applied to the distribution of research fronts and the small amount of residual articles. The latter are not part of research fronts because they are not bibliographically coupled, as none of their references are cited more than one time. We consider them as singular works that don't have any connections to other research fields. We compare the entropy $H = -\Sigma\, p_i \log p_i$, where $p_i = n_i / n$ and $n = \Sigma\, n_i$, to the maximum entropy $H_{\max} = \log n$, which corresponds to the case of n unclustered entities ($n_i = 1$, for all i).

We chose six countries – USA, Japan, France, United Kingdom, Germany and Russia – and calculated the contributions of these six countries to the whole set of electrochemical papers in 1998 and the contributions to the research fronts respectively. We assigned an article to a country according to the nationality of at least one of its authors. To these distributions forming the national research landscapes the entropy formula was applied again.

The processing of the data was done by means of perl scripts and in addition the network analysis tool Pajek was used for clustering and visualization of the cluster distributions.

3. *Results*

Table 2 shows the cluster distributions that form at different threshold levels. While the citation threshold remains stable at $c > 1$, the *Salton*-normalised co-citation threshold S is increased from $S > 0.1$ to $S > 0.9$.

For co-citation thresholds beneath 0.6 the single linkage clustering produces cluster distributions that are massively dominated by one macro cluster. However, we seek to determine a cluster distribution, which is comparatively the most appropriate one.

Between $S > 0.4$ and $S > 0.5$, there is a massive decrease in the amount of references that are clustered, whereas at the same time a lot of small clusters emerge. Due to the fact that 0.5 is the result of the normalization of the co-citation count 1 of two articles whose simple citation counts are both two – which is a very common combination – a lot of co-citation connections drop out at level > 0.5. So we choose the cluster distributions right before that peak (> 0.4 and ≥ 0.5) as a basis for calculating the research fronts. We also calculate the entropy for these cluster distributions, based on the total amount of all references that are cited at least twice, as indicator for the equivalent entropy values at the research fronts level. Fig. 1 shows entropy values of reference cluster distributions for different levels. Note the leaps between value 0.4 and 0.41 and between ≥ 0.5 (lower value) and > 0.5.

Figure 1: Entropy values of reference cluster distributions for different thresholds

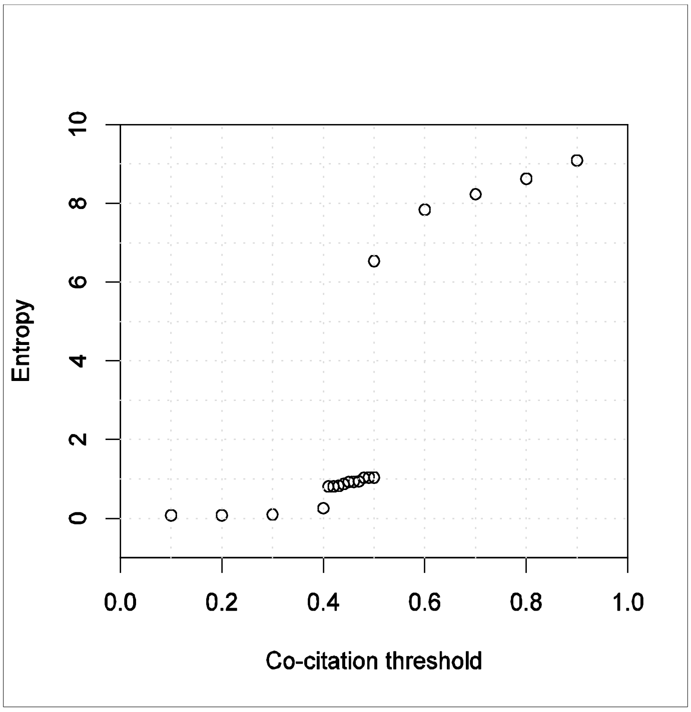
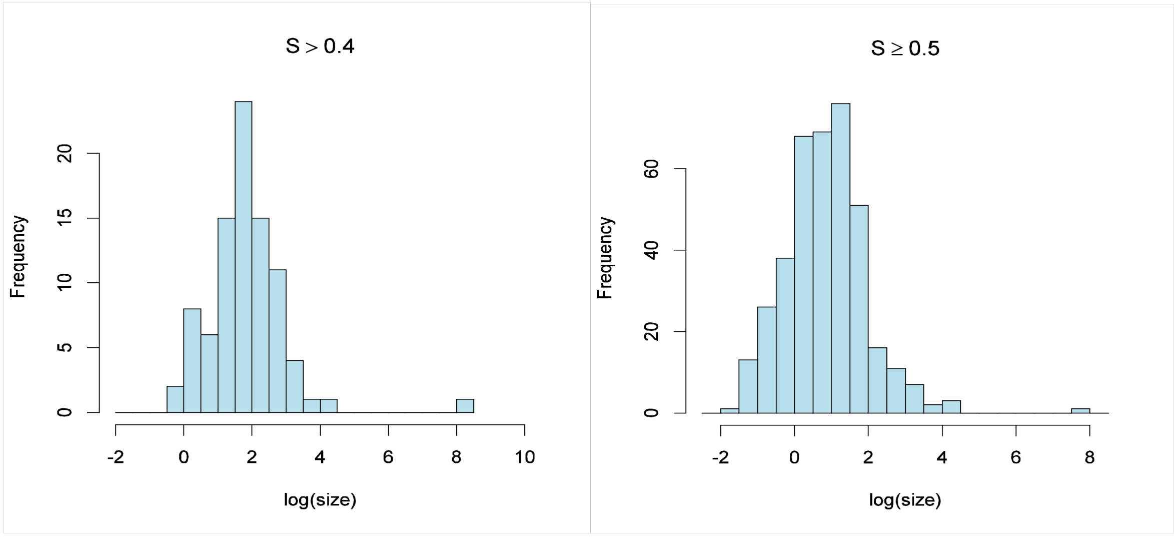


Table 2: Distributions of co-citation clusters

CO-CITATION LEVEL	NUMBER OF CLUSTERS	REFERENCES IN CLUSTERS	BIGGEST CLUSTER	SINGLE REFERENCES	H_c	$H_c/H_{c\ \max}$
> 0.1	18	12723	12 644	29	0.078	0.01
> 0.2	18	12723	12 644	29	0.078	0.01
> 0.3	20	12705	12 618	47	0.098	0.01
> 0.4	31	12544	12 421	208	0.256	0.03
≥ 0.5	126	11867	11413	885	1.037	0.11
> 0.5	990	8960	2630	3792	6.536	0.69
> 0.6	1262	8209	144	4543	7.841	0.83
> 0.7	1379	7385	73	1379	8.230	0.87
> 0.8	1435	6198	48	1435	8.626	0.91
> 0.9	1226	3869	25	1226	9.088	0.96

Figure 2: Size-frequency distribution of research fronts at the co-citation level $S > 0.4$ and $S \geq 0.5$, respectively.



For $S > 0.4$ we get 239 research fronts which comprise 3865 articles (89% of the total number of 4319 articles); the entropy of this distribution is 2.284 and $H_r/H_{r\,\max} = 0.32$.

For $S \geq 0.5$ the number of articles in research fronts is the same, but they are now dispersed over 1011 research fronts, $H = 4.096$ and $H_r/H_{r\,\max} = 0.49$.

In comparison to the cluster distributions the macro cluster that emerges at the level of research fronts is smaller in relation to the total number of clustered references. Whereas it then comprised nearly the whole number of clustered references, the macro cluster at the research fronts' level consists of 58 % (≥ 0.5) and 81% (> 0.4) respectively. Figures 2 and 3 show that – apart from the macro cluster – the research front log-size distribution is rather symmetric.

As there are a lot more research fronts at the level ≥ 0.5, the amount of very small entities is also larger, as Figure 2 shows. We consider the distribution to be too fragmented to give a plausible representation of a single field. So, we prefer to continue with the research fronts distribution at the threshold $S > 0.4$, to which the contributions of six countries are calculated.

Table 3: Distributions of research fronts of six countries

	ALL ARTICLES OF THE COUNTRY	NUMBER OF RESEARCH FRONTS	ARTICLES IN RESEARCH FRONTS	H_r
USA	763	114	670	2.061
Germany	393	80	345	1.901
Russia	320	48	271	1.899
Japan	658	105	597	1.866
France	356	81	333	1.737
UK	245	56	226	1.568

The USA being the country with the biggest output on research articles also has the most diverse research landscape according to both measures. The table shows that apart from the first rank, the two measures lead to different judgements of the diversities. For example, Japan has the second largest number of articles and ranks second according to its number of research fronts, but ranks only fourth in the diversity measurement with the Shannon index, while Russia has the smallest number of research fronts but ranks third according to the Shannon index

4. Discussion

This first attempt to measure the diversity of a scientific field has yielded both methodological and conceptual results. Firstly, several methodological points can be made. The single linkage cluster algorithm has proven to us to yield rather unsatisfying results. In all likelihood conditions the dominating macro cluster is a result of the chaining tendency of single linkage. As possible subsets that evolve from a common source cannot be identified by this algorithm as long as the source concept is still co-cited with more recent articles, single linkage clustering tends to produce large clusters that might conceal relevant substructures. In former studies that included the use of single linkage (Small & Sweeney [11]; van Raan [10], Jarneving [12]) maximal or minimal sizes are applied to level the extremely skewed distributions, but we consider this problematic as it is impossible to justify concrete size limitations on a theoretical basis.

The process of projecting the co-citation clusters to the current level actually flattens the distribution – the size of the macro cluster to the number of all clustered items is decreased in proportion to the cluster distribution – and by means of the additional application of bibliographic coupling the number of clusters is increased, too. But nevertheless, single linkage clustering should not be applied as a basic method, anymore.

We consider it to be a promising option to completely replace the co-citation method by bibliographic coupling. Jarneving [12] works out that structural differences scarcely exist between research fronts that are based on co-citation and those that are based on bibliographic coupling. Nevertheless, the process of projecting the co-citation clusters onto the current level imposes methodical difficulties in dealing with overlapping research fronts which can be circumvented by using bibliographic coupling as a single procedure.

As for alternative cluster algorithms we believe that the cluster definition of complete linkage is too strict to be feasible for citation data. An adequate model must imply uncitedness of relevant articles. The agglomerative-hierarchical average linkage can be applied well in combination with the cosine measure and is less strict than complete linkage, but more effective than single linkage. Radicchi et al. [13] and Newman [14] proposed interesting new graph analytical algorithms that are scalable especially for large datasets. Both are inspired by the Girvan-Newman-algorithm.

There is also possibly room for improvement in the constitution of the dataset. We could constitute the dataset by a search of title words of articles instead of first defining the journals that make up a field and then selecting all articles published in these journals. This approach may be more exact as according to *Bradford's Law of Scattering*, articles relevant to a certain field are not strictly confined to the journals of that field.

A second set of problems refers to the interpretation of the measures of diversity applied. The discrepancies between the two measures – number of research fronts and Shannon index – clearly demonstrate that great care is required in the definition and measurement of diversity. The number of research fronts is easier to interpret but neglects the distribution of research effort across research fronts. The Shannon index includes this information. However, the respective contributions of the number of research fronts and the distribution of articles across research fronts is difficult to evaluate.

Furthermore, diversity and size of a research profile do not seem to correlate directly when diversity is measured by the Shannon index, while the link is somewhat stronger when diversity is measured by the number of research fronts.

Finally, since the ultimate aim of measuring the diversity of research is to compare either the diversity of national fields or the diversity of one country's field at different points in time, it is essential to assess the differences between numbers of research fronts respectively H_i or $H(t_i)$. Table 3 cannot yet be interpreted in a policy context because it is not possible to establish which of the two measures (if any) is valid, and which of the differences are significant.

5. Conclusions

In order to measure the diversity of a country's research, fields must be delineated at the international and national levels, approaches within fields must be identified, and the distribution of research efforts across approaches must be measured. The feasibility study has demonstrated that a bibliometric approach to research diversity can solve these problems and is therefore a promising instrument for studies of scientific diversity. The limits of single linkage co-citation clustering that have been revealed warrant the search for alternative methods. Another problem that will return to the agenda when fields with less clear natural boundaries are going to be investigated is the methodology of field delineation. Finally, the study has demonstrated that a better understanding of measures of research diversity is necessary, which includes the test of other measures of diversity proposed in the biodiversity literature. Since the possibility of an unobtrusive objective measurement of diversity has been confirmed in principle, all these tasks appear to be worthwhile.

References

1. Adams, J. and D. Smith, Funding research diversity: The impact of further concentration on university research performance and regional research capacity, A report to Universities UK by Evidence Limited, 2003.
2. Harley, S. and F.S. Lee, Research Selectivity, Managerialism and the Academic Labor Process: The Future of Nonmainstream Economics in UK Universities. In *Human Relations*, 50, pages 1427-1460, 1997.
3. Grupp, H., The Concept of Entropy in Scientometrics and Innovation Research. In *Scientometrics* 18(3-4), pages 219-239, 1990
4. Magurran, A.E., *Measuring Biological Diversity*. Oxford, Blackwell, 2004
5. Gaston, K.J. and J.I. Spencer, *Biodiversity: An Introduction*. Oxford, Blackwell, 1998
6. Van Raan, A. F. J., Scientometrics: State-of-the-art. In *Scientometrics* 38, pages 205-218, 1997
7. Noyons, E.C.M. and A.F.J. van Raan, Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Research. In *Journal of the American Society for Information Science* 49, pages 68-81, 1998
8. Leydesdorff, L. and S. Bensman, Classification, Powerlaws, and the Logarithmic Transformation. In *Journal of the American Society of Information Science and Technology* (forthcoming), 2006
9. Ahlgren, P., B. Jarneving and R. Rousseau, Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. In *Journal of the American Society of Information Science and Technology* 54(6), pages 550-560, 2003
10. Van Raan, A. F. J., Fractal Geometry of Information Space as represented by Cocitation Clustering. In *Scientometrics* 20(3), pages 439-449, 1991
11. Small, H. and E. Sweeney, Clustering the *Science Citation Index* using Cocitations. In *Scientometrics* 7(3-6), pages 391-409, 1985
12. Jarneving, B., A Comparison of two Bibliometric Methods for Mapping of the Research Front. *Scientometrics* 65(2), pages 245-263, 2005
13. Radicchi, F., C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Defining and Identifying Communities in Networks. In *PNAS* 101(9), pages 2658-2663, 2004
14. Newman, M. E. J., Fast Algorithm for Detecting Community Structures in Networks. In *Physical Review* E 69, 066133, 2004