Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 66, No. 2 (2006) 365–376

How to identify research groups using publication analysis: an example in the field of nanotechnology

CLARA CALERO,[a] RENALD BUTER,[a] CECILIA CABELLO VALDÉS,[b] ED NOYONS[a]

[a] *Centre for Science and Technology Studies (CWTS), University of Leiden, Leiden (The Netherlands)*
[b] *Fundación Española para la Ciencia y Tecnología (FECYT), Madrid (Spain)*

We present a new bibliometric approach to identify research groups in a particular research field. With a combination of bibliometric mapping techniques and network analysis we identify and classify clusters of authors to represent research groups. In this paper we illustrate the application and potential of this approach and present two types of outcomes: actual research groups and potential research groups. The former enables us to define research groups beyond the organizational structure. The latter may be used to identify potential partners for collaboration. Our approach is a starting point to deal with the complex issue of research groups in a changing structure of scientific research.

Introduction

On a conceptual level, much has been made of the observed switch from "Mode 1" to "Mode 2" models or types of research and knowledge generation put forward by Michael Gibbons and co-workers (GIBBONS et al., 1994). The model shift has been related with the trend towards multi- and inter-disciplinary research and the long term decline of single discipline research, but also in the increased wish and need for collaboration in researchers. (PREST, 2000). This approach recognizes that research is a collective effort, combining diverse actors, competences and capabilities. It puts the emphasis upon the collective setting, intermediary between individual researchers and research institutions (LAREDO, 2003). Although the typical research group still has a core team consisting of tenured staff and students (graduate, doctoral and postdoctoral), there is usually a more peripheral level of visiting scientists and cooperating domestic and foreign colleagues. And actually are these broad cooperative elements the actual research-performing units, which may reflect the realities of the scientific process more accurately that do core teams. (SEGLEN & AKSNESS, 2000). In such a framework, policies/strategies cannot rely only on a content dimensions i.e., thematic priorities, they have also to care about organizational aspects. Questions such as: Do we have the right

Received June 22, 2005

Address for correspondence:
CLARA CALERO
Centre for Science and Technology Studies (CWTS), University of Leiden
Wassenaarseweg 52, P. O. Box 9555, 2300 RB Leiden, The Netherlands
E-mail: clara@cwts.nl

0138–9130/US $ 20.00

research groups? Are they inter-connected enough? What about their connections with their environment?, are more and more pressing (LAREDO, 2003).

It has been stated often that bibliometric analyses could play an important role in measure these tendencies and along this road a number of studies have been carrying out lately. Recently many improvements have been made in getting an overview of multi-and inter-disciplinary fields through bibliometric maps (NOYONS, 1999; NOYONS et al., 2002; NOYONS et al., 2003). The co-authorship data were used in many studies to measure collaboration (e.g., PERSSON & BECKMANN, 1995; MELIN & PERSSON, 1996; BORDONS & GOMEZ, 2000; SEGLEN & AKSNESS, 2000). And starting around 2000 several researchers began the construction of large-scale networks using co-authorship data representing research in mathematics (BARABÁSI et al., 2002); biology, physics and computer science (NEWMAN. 2001); and neuroscience (BARABÁSI et al., 2002). However, most of these studies are fragmented, focusing on one or a few characteristics of the process at a time. Only a few attempts have been made to relate cognitive structures and collaboration (e.g., MUTSCHE & QUAN HAASE, 2001). Here, we used bibliometric mapping techniques and network analysis to identify and classify research groups. This approach intends to cover the two key trends of the knowledge process: multi-and inter-disciplinary scientific fields and broad cooperative units of research.

The aim of this paper is to present a new approach to identify research groups analyzing the articles published in scientific journals in a particular science field.

Data and methods

The data for this study were taken from a project financed by the Spanish Foundation for Science and Technology (FECYT). One of the objectives of the project was to map and identify Spanish research groups in the field of nanotechnology.

Delineation (publication data collection)

The data collection (or delineation) procedure was carefully designed in close collaboration with the field experts. In a first step, core publications for the field were collected. The database Current Contents from the Institute for Scientific Information (ISI) was used as a source of primary data. This primary collection was based on the delineation adopted in the EC mapping of excellence project (NOYONS et al., 2003). We took the final discussion from that project into consideration and compiled a primary search strategy for the FECYT project. From this core set of publications we extracted candidate search terms to expand the set of publications and asked experts to indicate the relevant candidates (or suggest alternatives). The FECYT and the expert groups involved in this project considered that for the delineation of this field special emphasis

should give to materials, because the importance they have in Spain. In a second round the suggested terms were used and a new data were collected. The results in this study were based on this second search strategy.

The core publications for the field were collected by the following search terms:

• nano* NOT (nanomet* OR nano2 OR nano3 OR nano4 OR nano5 OR nanosecon* OR nano secon*) OR
• nanomet* scale* OR nanometerscale* OR nanometer length OR nano meter length
• nanoa* OR nanob* OR nanoc* OR nanod* OR nanoe* OR nanof* OR nanog* OR nanoh* OR nanoi OR nanoj* OR nanok* OR nanol* OR nanon* OR nanoo* OR nanop* OR nanoq* OR nanor* OR nanot* OR nanou* OR nanov* OR nanow* OR nanox* OR nanoy* OR nanoz*
• atom* force microscop*
• tunnel* microscop*
• scanning probe microscopy
• scanning force microscop*
• semiconductor quantum dot
• silicon quantum dot
• quantum dot array
• coulomb blockade
• Single molecule
• molecular motor
• molecular beacon
• biosensor
• self-organized growth
• electron beam lithography
• monolayers growth
• optoelectronic* device*
• Quantum Computing
• quantum devices
• quantum Discs
• quantum optoelectronics
• quantum Wells
• quantum wires
• Scanning probes techniques
• Transmission electron microscopy
• resonant cavity
• resonant cavities
• self assembling
• self ordering
• spintronics
• submicron devices
• vertical cavity surface emitting Laser*
• cantilevers
• quantum dots
• Molecular Beam Epitaxy

The final set from the period January 1996–January 2003 contained a total of 91,372 articles retrieved from the above mentioned database Current Contents.

Bibliometric mapping

The bibliometric map is a two-dimensional representation of the core publications collection, designating the field of nanotechnology. From these publications, we extracted noun phrases from titles and abstracts to be used for a co-occurrence analysis. This co-occurrence analysis clusters selected noun phrases (keywords). These keywords were identified from the endless list of noun phrases, on the basis of bibliometric distribution, syntactic features and (semantic) content.

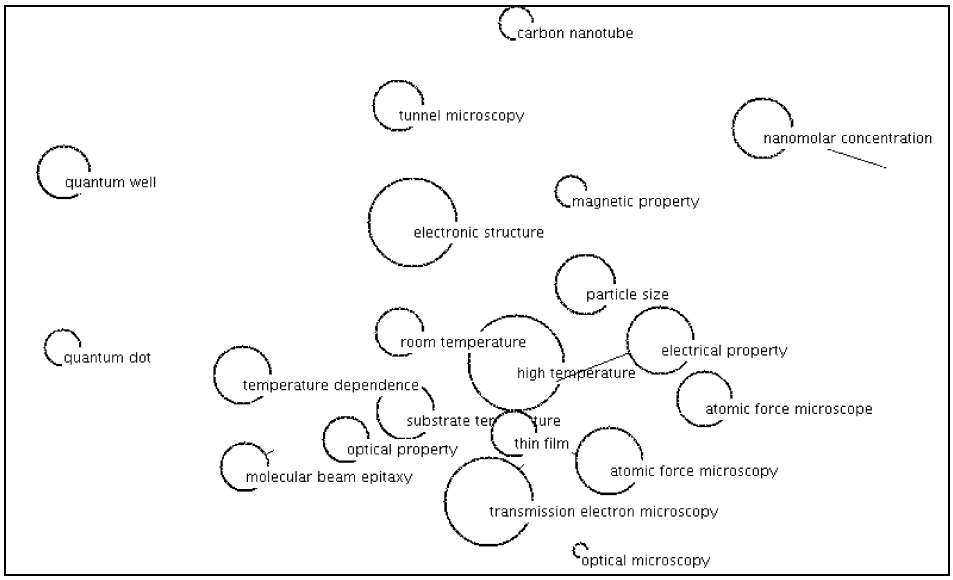

Figure 1. Bibliometric map of nanotechnology
The sub-domains (clusters of topics) are positioned, depending on the cognitive orientation. The more two sub-domains are related the closer they are. Each sub-domain is characterized by the most prominent keyword within. The size of the surface indicates the number of publications represented.

The clusters of keywords designated sub-domains in the field. By these keywords, publications were assigned to sub-domains. Thus, sub-domains were in fact, subsets of publications from the entire collection, the field. As publications might be assigned to more than one sub-domain, we generated a co-occurrence matrix of sub-domains. The cells in this N*N matrix (in which N designates the number of sub-domains), contained

the number of publications overlapping in two of the N sub-domains. This matrix was the input for Multidimensional scaling (MDS). This technique put the N elements in a two-dimensional space in such a way that sub-domains with a similar orientation in relation to all other the sub-domains, were in each other vicinity, whereas sub-domains with a different orientation were distant from each other. This two dimensional representation was the bibliometric map of the field.

Figure 1 shows the bibliometric map of the field of nanotechnology for the present study.

Identification of a research group

The analysis was based on units formed by combinations of author name and main organization. In this case, because of the scope of the project, all the organizations selected are from Spain. The research groups were identified and defined on the basis of similar research activity profiles and co-authorship.

Author/Organization Combination (AOC). We assumed that we could define a group bibliometrically by a collection of publications. This collection was identified by the oeuvres of one or a set of authors. In order to do that, we had to deal with the publication author names. We encountered two problems in publications data related with the author's field: two persons with the same author name (homonymous names) or two or more author names referring to the same person (synonymous names).

To solve the problem with homonymous names, we used a combination of author names and main organization (university, company…). Each publication had in most cases at least one author and at least one address (from the address field it was considered just the organization); in this case the only thing we knew was that the first author is attached to the first organization. But as the first author may also be at the second or third organization and the second author may be attached to any of the organizations… We assigned in a publication all author names to all organizations. So for a publication with 3 authors and 2 organizations,[1] we defined in fact 6 (3×2) authors. Or more correctly, we define 3 authors associated with two organizations.

[1] Actually in the publication itself the author name is attached to the organization/s that belongs to. This doesn't happen on the information contained on the electronic databases.

Example of AOCs

Publication X
Authors
1 A 1
2 A 2
3 A 3
Addresses
- Org A
- Org B

=>

3×2

AOCs in Publication X
- A 1, Org A
- A 2, Org A
- A 3, Org A
- A 1, Org B
- A 2, Org B
- A 3, Org B

This solution, however, increased the problem of the second type (synonymy). As each author was associated with all the organizations in a publication, more names were referring to the same person. The analysis of the relations between the AOCs dealt with this problem.

Because the purpose of this study was to identify Spanish research groups, we selected only the organizations coming from Spain. Besides, only AOCs with more than six publications were considered.

Activity similarity relations. Using the bibliometric mapping and clustering analysis of the field of nanotechnology we created the research profile of each AOC. In our database this profile was compiled by the number of publications the AOC had in each cluster.

The next step was to compare the AOCs on the basics of this activity profile. In order to do so we used a similarity measure, the cosine coefficient (NOYONS, 1999). Each pair of AOCs got a value between 0 and 1 indicating their similarity. Because the objective of our study was to identify research groups based on their research activity similarity, we considered only the relations with a cosine coefficient higher than 0.9.[2]

Co-publication relations. We constructed the co-author/organization matrix composed by co-occurrences of the AOCs co-publishing the same article.

Network analysis. A network analysis was applied to represent the two different relations explained above and to identify community structures. In the network theory the graphs are composed of nodes (or actors or points or vertices) connected by edges (or relations or ties).

[2] The threshold of 0.9 was arbitrary, but a small test using other thresholds did not yield significantly different results.

For the identification of subgroups of authors within the network, we used the k-core approach. A k-core is a subgraph in which each node (AOC) is connected to at least a minimum fixed number (K) of the other nodes in the subgraph. The k-core approach is less strict (compared with others like cliques, n-cliques, n-clans…), allowing actors to join the group if they are connected to k members, regardless of how many other members they may not be connected to. (WASSERMAN & FAUST, 1994).

So the Activity Similarity Graph represented the relations between AOCs connected by similar research activity profile, while the Co-author Graphs represented the relations between AOCs by the absolute number of co-publications.

The Activity Similarity Graph was used as a base for the analysis. We extracted the subgraphs from this network. Each subgraphs extracted was analyzed also using the co-authorship.

Results

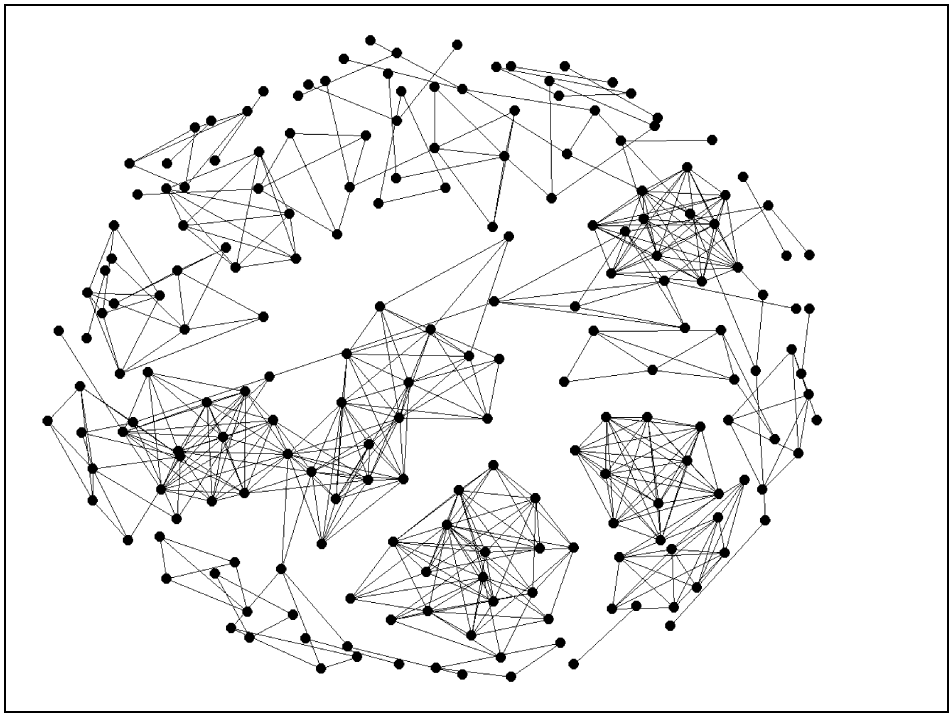

Figure 2. Activity Similarity Graph of nanotechnology
Each node represents an author/organization combination (AOC).
A connecting line indicates a similar research profile.

The results expose in this section are two examples of the subgraphs extracted from the activity similarity graph and illustrate the application and potential of this new methodology. These cases are representative of the two types of outcomes expected from this method: the identification of a research group and the detection of potential partners.

In the Activity Similarity Graph each node represents an AOC and the relations depict a similar research profile (Figure 2). This graph is the starting point to identify subgroups of AOCs. As mentioned in the previous section we used the K-core approach to divide this network into subgraphs.

Identification of a research group

Figure 3 shows a subgraph of ten AOCs; each of them is related with, at least, eight of the others. We have identified already a group of AOCs with a very similar research profile. Consequently this is a potential research group, but are these AOCs working together? Figure 4 illustrates the connections between these ten AOCs based on their co-authorship. As we can see these AOCs are actually working together. So they are a research group. The last step it will be to assign the AOCs to the authors and organizations that they are related with.

If we take a closer look to the individual AOCs, we can see that for each author name there are two organizations related: Cádiz University and Polytechnic University of Madrid (UPM). But we are not concern about an author or an organization we are looking for a group. With the information contained in Figures 3 and 4, we have identified a research group composed of six researchers: Izpura I, Gutiérrez M , Aragón G, González D, García R, and Sánchez JJ; coming from two universities: Cádiz University and Polytechnic University of Madrid.

Identification of potential research partners

The second example illustrates an other sub-graph extracted from the Activity Similarity Graph (Figure 5). In this case each AOC is connected with at least three of the others. As the previous example, the relations mean a similar research profile. Again, we can take a look at Figure 6 to the co-authorship relations. Here, we get two co-publications groups. At the top part of Figure 6, there is a co-author group coming from Alicante University formed by four researchers: Herrero E, Rodes A, Feliú JM, and Gomez R. The second co-author group belongs to the Autonomous University of Madrid (UAM) and to the CSIC-ICMA (one of the institutes at the Spanish Council for Scientific Research) formed by three researchers: Gómez Herrero J, Ordejón P and Baró AM. Finally, there are two AOCs not co-publishing in this set: Agrait N/UAM and Flores F/UAM.

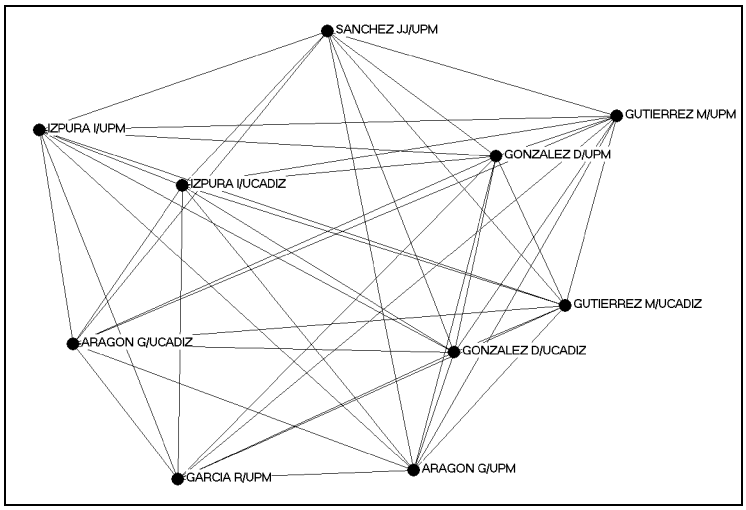

Figure 3. 8-core subgraph based on activity similarity relations
Each pair of AOCS connected represents a similar research profile.

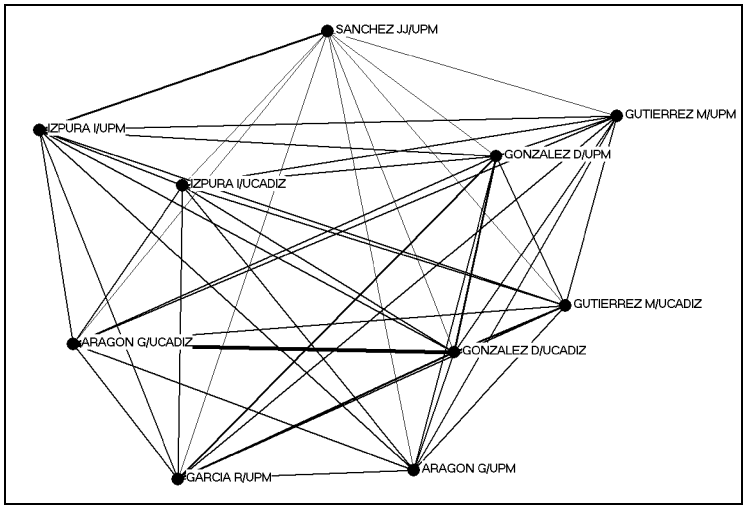

Figure 4. Subgraph based on co-authorship relations
Each pair of AOCS connected shows a co-publishing activity.

The activity similarity subgraph has identified a group of AOCs with similar research profile. While the coauthor data depicts that this set is divided into two co-authors groups and two isolated AOCs. The information provided by the activity similarity subgraph identified potential partners for the AOCs that not co-publish.

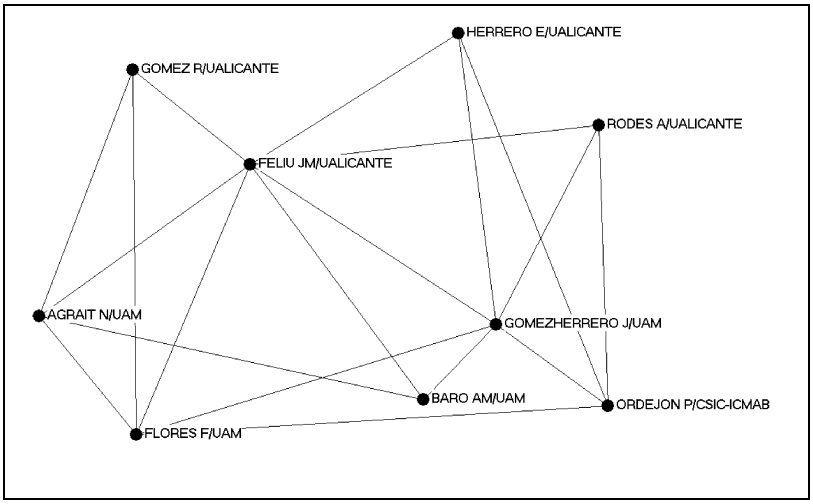

Figure 5. 3-core subgraph based on activity similarity relations
Each pair of AOCS connected represents a similar research profile.

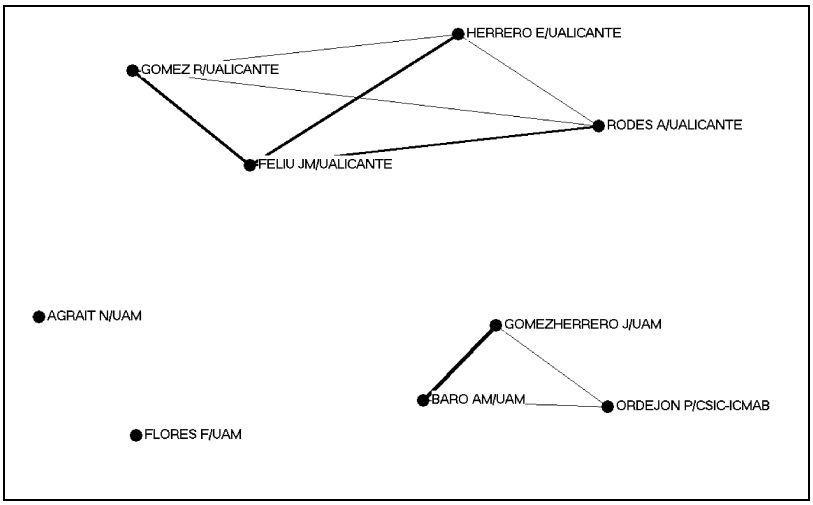

Figure 6. Subgraph based on co-authorship relations
Each pair of AOCS connected shows a co-publishing activity.

Conclusions and discussion

In the last years we have observed considerable advances in measuring the knowledge production and utilization. The idea that the scientific research is moving from a personal, disciplinary-based, and place-bound ideal towards a collective, problem-oriented and multi-organizational activity is well-accepted nowadays.

The method presented here should be considered only as the starting point toward a complete methodology for identifying research groups and potential research partners in scientific fields. A first and important result of the study regards with the matter that we have identified functional rather than physical groups. Following SEGLEN & AKSNES (2000) definition of a research group: "…a research group assignment based on co-authorship defines functional rather than physical groups, and might include, e.g. authors with whom a group member has collaborated in connection with a short-term scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". The groups are defined over a six year time period meaning that the group members have not necessarily worked together. In addition, the identification of the members through the AOCs allows the same person to belong to more than one group. This is the case, for instance, of a researcher that moves from organization and changes his line of research.

A second significant outcome of the study concerns the idea of being able to identify potential research partners. Using the activity similarity relations combined with the co-author relations it is possible to detect groups working on the same areas but not co-publishing.

A third important result of our approach is that we should be able to deal with the homonymous and synonymous names. The combination of the author and the address field in a publication allow us to solve the problem of the homonymous names, while the network analysis provides a possibility to deal with the latter. The combined data enables us to assign more accurately author names to authors.

Nevertheless, we are aware that there are a number of potential improvements that can be made to the method presented here, including the following suggestions to be implemented in further research:

- Validate the results with the opinions of the experts in the field.
- Analyze in more detail the profile and position of some authors in the activity similarity network to identify authors that are 'bridges' between groups with different profiles.
- Add to the analysis the impact factor for each AOC to identify success teams.
- Use other techniques related with community structures in networks and compare with the results from the K-core approach.
- Enlarge the scope of the analysis to international collaborations.
- Another issue to consider is the time evolution of the identified groups.

In summary, the method and results presented here should be considered a starting point for developing a methodology to identify systematic research groups. It is important to note that such method is open: more details are going to be incorporated which are going to improve the results.

References

BARABÁSI, A.-L., JEONG, H., RAVASZ, E., NÉDA, Z., SCHUBERT, A., VICSEK, T. (2002), Evolution of the social network of scientific collaborations, *Physica A,* 311 : 590–614.

BORDONS, M., GÓMEZ, I. (2000), Collaboration networks in science. In: H. B. ATKINS, B. CRONIN (Eds), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield, Information Today*, Medford, NJ, pp. 197–213.

GIBBONS, M., LIMOGES C., NOWOTNY, A., SCHWARTZMAN, S., SCOTT P., TROW, M. (1994), *The New Production of Knowledge: The Dynamics of Science and research in Contemporary Societies*, Sage, London.

LAREDO, P. (2003), University research activities. On-going transformations and new challenges, *Higher Education Management and Policy*, 15 (1) : 138–163.

MELIN, G., PERSSON, O. (1996), Studying research collaboration using co-authorships, *Scientometrics*, 36 : 363–377.

MUSTCHKE, P., Haase, A. Q. (2001), Collaboration and cognitive structures in social science research fields. Towards socio-cognitive analysis in information systems, *Scientometrics*, 52 (3) : 487–502.

NEWMAN, M. E. J. (2001), The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci, USA,* 98 : 404–409.

NOYONS, E. C. M. (1999), *Bibliometric Mapping as a Science Policy and Research Management Tool,* Thesis Leiden University. Leiden: DSWO Press (ISBN 90-6695-152-4).

NOYONS, E. C. M., Buter, R. K., VAN RAAN, A. F. J. (2000), *Mapping the Field of Neuroscience.* Electronic version with interactive facilities available via www.cwts.leidenuniv.nl

NOYONS, E. C. M.,. BUTER, R. K., VAN RAAN, A. F. J., SCHMOCH, U., HEINZE, T., HINZE, S., RANGNOW, R. (2003), *Mapping Excellence in Science and Technology across Europe: Nanoscience and Nanotechnology.* Report of project EC-PPN CT 2002-0001 to the European Commission, Leiden.

PERSSON, O., BECKMAN, M. (1995), Locating the network of interacting authors in scientific specialties, *Scientometrics,* 33 : 351–366.

PREST (2000), *Impact of the Research Assessment Exercise and the Future of Quality Assurance in the Light of Changes in the Research Landscape,* prepared for Higher Education Funding Council for England (HEFCE), April 2000, available from www.hefce.ac.uk

SEGLEN, P. O., AKSNES, D. W. (2000), Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, *Scientometrics,* 49 : 125–143.

WASSERMAN, S., FAUST, K. (1994), *Social Network Analysis*, Cambridge University Press, Cambridge.