Detecting Research Groups in Coauthorship Networks

Antonio Perianes-Rodríguez[1,4] Carlos Olmeda-Gómez[2,4] Félix Moya-Anegón[3,4]

05 June 2008

Abstract

From the perspective of Library Science and Information Science, little research has yet been conducted on scientific networking and its possible uses in ascertaining the composition of research groups, the differences in associations between specialities or departments, and the different policies that may be followed in this regard, depending on the institution or the domain analyzed. Traditionally, most studies on scientific collaboration have been geared to analyzing output, be it international or domestic, of a given scientific discipline or a research institution. Studies on smaller units such as departments or research groups are less common.

This work focuses on a specific facet of scientific communication networks, namely scientific co-authorship networks, based on the premise that scientific communication is the essence of research, and research is only known as such when it has been analyzed and accepted by the scientific community, which gives it the status of a social activity. The use of the term "scientific communication", therefore, means deliberately limiting considerations on communication to a specific group of individuals (authors directly involved in the creation of original research work): those engaging in a well-defined activity and having very specific objectives.

The main objective of this work is to identify, characterize and interpret research groups in Carlos III University of Madrid using empirical analysis, through the examination and visualization of scientific networking based on co-authorship papers. The findings obtained will contribute to a better understanding of network dynamics and of how they affect network topology and the internal structure of links among such research groups, and by extension, how they affect the higher-level administrative units of which they form a part. To this end, this work will try to achieve two specific objectives: on one hand, to model and characterize co-authorship networks by calculating indicators of the properties of nodes and links that describe sizes and neighbourhoods in subgraphs, as well as to obtain comprehensive measurements that statistically characterize the structure of network interconnections as a whole. On the other hand, to create specialized network-based visualizations, including diagrams of nodes and links, that can be used as interfaces to retrieve information. These interfaces provide data on the element matrices and on the values of their attributes in a clear, easily understood, explanatory and interactive way. They facilitate an understanding of the structural context represented, transmitting detailed information to the user about a variety of aspects relating to scientific collaboration and its evolution over time, such as administrative position, gender, speciality areas of research and the internal and external association patterns among authors.

1 Introduction

The teamwork intrinsic in scientific activity from the very dawn of science is still characteristic of research today, as evinced by the growing specialization and internationalization that has taken place in recent decades (Beaver D and

[1] Carlos III University of Madrid, Department of Library and Information Science, Getafe, Spain, aperiane at bib dot uc3m dot es
[2] s. footnote before, carlos dot olmeda at uc3m dot es
[3] University of Granada, Faculty of Library and Information Science, Granada, Spain, felix at ugr dot es
[4] SCImago Research Group, www dot scimago dot es

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

Rosen R, 1978), (Harsanyi MA, 1993), (Melin G and Persson O, 1996).

Any number of bibliometric and scientometric studies stand as proof of the growing interest in the microanalysis of research activity, focusing on the research group level (Bordons M and Zulueta MA, 1997). This has been largely due to the fact that guidelines for the concentration of research are among the priorities of scientific policy, in particular measures relating to research management and organization. But it is also because research groups are the basic organizational unit for universities aiming to assume, organize around and acquire a business approach and steer their technological and research activities toward the establishment of links with the surrounding business community (Etzkowitz H, 2003).

Although studies on the activity, composition and productivity of research groups through micro-level bibliometric indicators provide insight into research structure and dynamics (Hou H et al., 2008) (von Tunzelmann N et al., 2003), such analyses have seldom been conducted. This is due to the existence of many types of technical or even technological difficulties, including the application of statistical methods to small quantities or the costs of gathering and processing duly normalized data for detailed analysis involving such minute disaggregation.

The present paper proposes a method for detecting, identifying and visualizing research structures. The aim pursued is to contribute to the microanalysis of internal research dynamics at the individual and research group level, based on scientific co-authorship networking by the members of Carlos III University of Madrid (UC3M) departments.

2 Method

2.1 Definition of a Research Group

Many attempts have been made to formalize what is meant by research group, one of the most precise definitions being a community of scientists who work together in the approach to and development of research, sharing material and financial resources, but not necessarily organized along the lines of the formal structure of the institution or institutions where the activity is conducted (Zulueta MA and Bordons M, 1999).

The proposals put forward by the various areas of science to delimit the data identifying research groups refer to such delimitation in different ways. Cohen identified two methodological patterns to delimit groups: result-based and input-based (Cohen JE, 1991). Under the former, researchers in the same department, research partners or co-authors of scientific papers, regardless of their affiliation, are regarded to be members of a research group. In this case the research population is defined on the grounds of co-author details or citations (Noyons ECM et al., 1999). Productivity studies based on bibliometric techniques constitute an example of this approach, in which teams are represented by author networks deduced from the frequency of co-authorship. Groups are not necessarily administrative or institutional units. On the contrary, such analyses identify operational rather than physical groups (Seglen PO and Aksnes DW, 2000). This precludes the need for prior information on the unit to be studied. In input-based method, by contrast, author affiliation is required to be able to conduct the analysis. The result-based method, however, omits scientists who do not publish, whereas input-based studies define groups on the basis of administrative agreements that include all members, whether or not they publish. This paper uses a combination of the two patterns defined by Cohen to obtain networks apt for detecting and classifying research groups.

Groups, then, are subsets of closely related nodes on a graph. Analysis of nodal groups is a valuable tool for understanding networks. Such analysis entails essentially two tasks: detection and identification. The former consists in discovering the different groups existing in the network, while the latter focuses on characterizing each subset of nodes extracted from the initial network.

Many detection algorithms use hierarchical clustering techniques. Such algorithms are two-phased: the first defines the metrics that represent internode similarity. The second uses extraction methods defined on the basis of two possible types of metrics (Balakrishnan H and Deo N, 2006), agglomerative and divisive (Donetti L and Muñoz MA, 2004), (Newman

MEJ, 2004a), (Newman MEJ, 2004b), (Radicchi F et al., 2004), (Reichardt J and Bornholdt S, 2004), (Wu F and Huberman BA, 2004).

But in addition to hierarchical extraction methods, which are only useful if the structure is to be interpreted in terms of sets of separate communities, others based on locating network communities by statistical analysis of the raw data are also available (Palla G et al., 2005). Of the many such schemes in place, the one chosen for the present study is factor analysis, which has been widely used in similar analyses (Chen C et al., 2001), (Chen C and Carr L, 1999a), (Chen C and Carr L, 1999b), (Chen C and Paul RJ, 2001), (Ding Y et al., 2000), (McCain KW, 1990), (White HD and McCain KW, 1997). Its use is justified in that groups can be defined on the grounds of the structure of interconnections or, in other words, the premise that the members of each group tend to choose and by chosen by the same partners. Consequently, membership in a group is established on the basis of similarities between the choices made by and about each author. Such conditions make choices tend to exhibit reciprocity, while the factors obtained and rotated form a simple structure. In short, the notion of group proposed is not a single common space in which all the participants are interlinked. Rather, the members of each group share a distinctive perceptive structure with respect to their work, matching a different dimension in factor space in each of the resulting communities.

Finally, since the groups can be identified and aggregated in terms of common characteristics, the JCR subject categories corresponding to the bibliographic references cited in the papers by the authors constituting the factor were taken into consideration when assigning the name that characterizes each factor (Moya-Anegón F et al., 1998), (Vargas-Quesada B and Moya-Anegón F, 2007). Be it said in this regard that when a given characteristic (working in the same speciality, in this case) is relevant to the choice of authorship, the existence of two or more groups related to that common characteristic, but in different factors, cannot be ruled out.

3 Data

A relational database built with records for the period 1990-2004 taken from the Web of Science (SCI-expanded, SSCI and A&HCI), in which at least one author was affiliated with the UC3M, was used for the bibliometric analysis of the research conducted in the institution. The Institute for Scientific Information (presently Thomson Scientific) assigns each journal one or several subject categories. Journal Citation Reports (JCR) for both Science and Social Science for the years analyzed was the reference used to assign each paper a subject (ISI category).

3.1 Data Refinement

Once the assumption that a group can be defined from a collection of published papers signed by a series of authors is adopted, it necessarily follows that their names must be standardized and processed for that purpose (Calero C et al., 2006). Two problems are commonly encountered in connection with the author field: homonymy (two authors with the same name) and synonymy (the existence of different variations on an author's name). To obviate these difficulties, the SCImago group used ad hoc software that avoids homonymy by combining author and institution and synonymy by combining author and paper (Gálvez C and Moya-Anegón F, 2006), (Gálvez C and Moya Anegón F, 2007).

4 Results

The formulation of co-authorship networks under the premises described in the section on methodology provides fuller insight into the evolution of collaboration in each of the units analyzed.

By incorporating input-based data, the use of colour to differentiate actors in the visual information supplements the result-based network. Likewise, further to the premises introduced by Moreno, the use of variation in node size facilitates the visualization of each actor's characteristic features and the position of the vertices (Moreno JL, 1953).

The application of factor analysis to the network matrix identified underlying factorial groups on the grounds of the structure of their choice of bonds.

As noted in the description of the methodology, the JCR categories for the bibliographic references cited by the authors constituting the factors were used to interpret the results. Although common characteristics can be attributed to the factorial groups in a number of ways and the research speciality classification chosen is coarse-grained, the advantage of this approach is that it is inherent in the data themselves.

5 Discussion

A functional representation of research groups was obtained from co-authorship-based links through a suitable combination of result-based studies and administrative information from the university itself (input).

The resulting groups showed not only individual relationships, but how these relationships are able to draw authors together in larger structures, revealing the social and intellectual ties in the form of components whose isolation is an initial approximation of the concept of research group.

The method proposed proves to be useful and valid for detecting and identifying research groups defined on the basis of the structure of the choice of co-authorship bonds, with no need to disassemble the resulting networks or isolate any of their components. The result is a series of communities, each of which shares guidelines and objectives distinctly different from those of all the others, based on their research work.

This paper raises new challenges for the analysis of the properties of co-authorship networks, such as the observation of their organizational forms; the nature of their information flows; the prominence of and interaction between the actors; internal group functions within the complex system of which they form part, deduced from a combination of bibliometric and structural indicators; and the evolution of networks and groups that shed light on their respective "life cycles" by including information over time to observe their birth, transformation through aggregation or segregation and, as appropriate, disappearance. Lastly, the extension and comparison of this methodology to higher order aggregations (scientific, regional, national and international domains) is a very promising line of research.

References

Balakrishnan H; Deo N (2006). Discovering communities in complex networks. *Proceedings of the Annual Southeast Regional Conference, 44*. New York: Association for Computing Machinery. P. 280-285.

Beaver D; Rosen R (1978). Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. *Scientometrics*, 1, p. 65-84.

Bordons M; Zulueta MA (1997). Comparison of research team activity in two biomedical fields. *Scientometrics*, 40 (3), p. 423-436.

Calero C; Buter R; Cabello C; Noyons ECM (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics*, 66 (2), p. 365-376.

Chen C; Carr L (1999a). Visualizing the evolution of a subject domain: a case study. *IEEE Visualization*. San Francisco: IEEE Computer Society. P. 449-452.

Chen C; Carr L (1999b). A semantic-centric approach to information visualization. *International Conference on Information Visualisation, 3*. London: IEEE Computer Society. P. 18-23.

Chen C; Paul RJ (2001). Visualizing a knowledge domain's intellectual structure. *IEEE Computer*, 34 (3), p. 65-71.

Chen C; Paul RJ; O'Keefe B (2001). Fitting the jigsaw of citation: information visualization in domain analysis. *Journal of the American Society for Information Science and Technology*, 52 (4), p. 315-330.

Cohen JE (1991). Size, age and productivity of scientific and technical research groups. *Scientometrics*, 20 (3), p. 395-416.

Ding Y; Chowdhury GG; Foo S (2000). Journal as markers of intellectual space: journal co-citation analysis of information retrieval area, 1987-1997. *Scientometrics*, 47 (1), p. 55-73.

Donetti L; Muñoz MA (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical*

Mechanics: Theory and Experiment, 10012, p. 1-15.

Etzkowitz H (2003). Research groups as quasy-firms: the invention of the entrepreneurial university. *Research Policy*, 32 (1), p. 109-121.

Gálvez C; Moya-Anegón F (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics*, 69 (2), p. 323-345.

Gálvez C; Moya-Anegón F (2007). Standardizing formats of corporate source data. *Scientometrics*, 70 (1), p. 3-26.

Harsanyi MA (1993). Multiple authors, multiple problems. Bibliometrics and the study of scholarly collaboration: a literature review. *Library and Information Science Researc*h, 15, p. 325-354.

Hou H; Kretschmer H; Liu Z (2008). The structure of scientific collaboration networks in Scientometrics. *Scientometrics*, 75 (2), p. 189-202.

McCain KW (1990). Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science*, 41 (6), p. 433-443.

Melin G; Persson O (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36 (3), p. 363-377.

Moreno JL (1953). *Who shall survive? Foundations of sociometry, group psychotherapy and sociodrama*. New York: Beacon House.

Moya-Anegón F; Jiménez Contreras E; Moneda Corrochano M (1998). Research fronts in library and information science in Spain. *Scientometrics*, 42 (2), p. 229-246.

Newman MEJ (2004a). Detecting community structure in networks. *European Physical Journal B*, 38, p. 321-330.

Newman MEJ (2004b). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (suppl. 1), p. 5200-5205.

Noyons ECM; Moed HF; van Raan AFJ (1999). Integrating research performance analysis and science mapping. *Scientometrics*, 46 (3), p. 591-604.

Palla G; Derényi I; Farkas I; Vicsek T 2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435, p. 814-818.

Radicchi F; Castellano C; Cecconi F; Loreto V; Parisi D (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (9), p. 2658-2663.

Reichardt J; Bornholdt S (2004). Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93 (21), p. 218701-1-218701-4.

Seglen PO; Aksnes DW (2000). Scientific productivity and group size: a bibliometric analysis of Norwegian microbiological research. *Scientometrics*, 49 (1), p. 125-143.

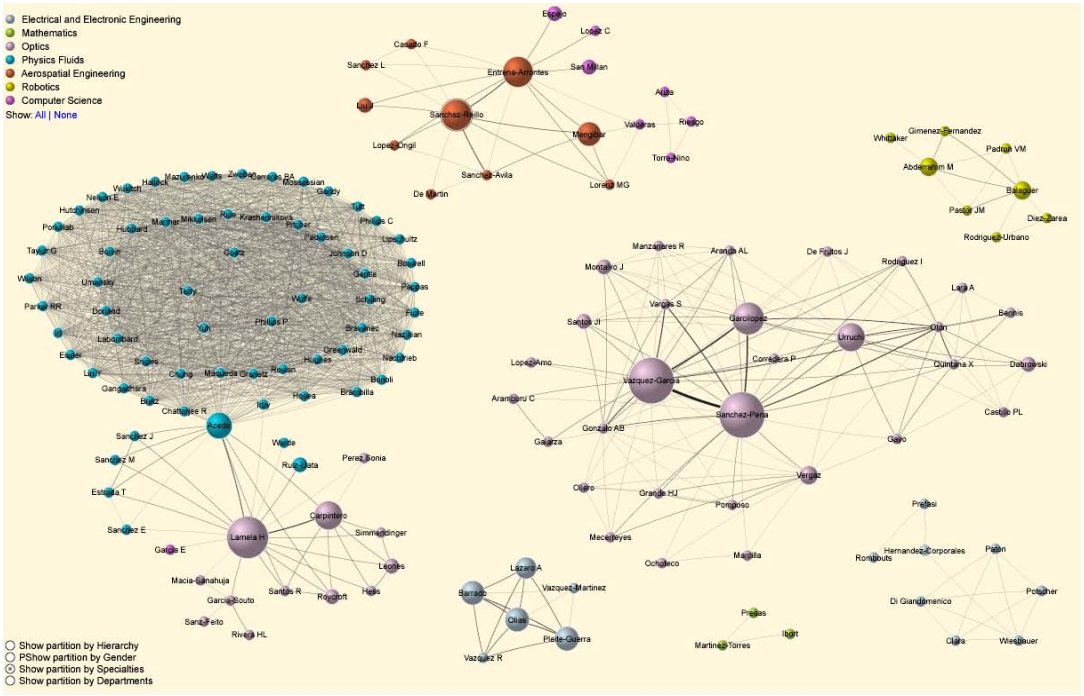
Vargas-Quesada B; Moya-Anegón F (2007). *Visualizing the structure of science*. Berlin: Springer.

von Tunzelmann N; Ranga M; Martin BR; Geuna A (2003). *The effects of size on research performance: a SPRU review*. Brighton: University of Sussex.

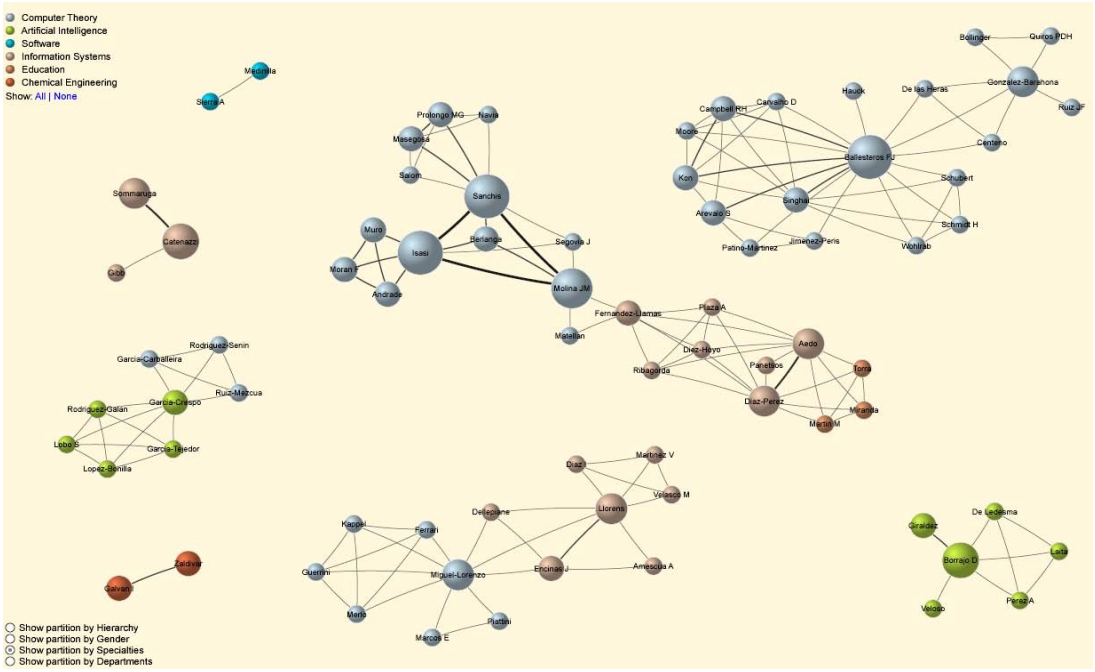
White HD; McCain KW (1997). Visualization of literatures. *Annual Review of Information Science and Technology*, 32, p. 99-168.

Wu F; Huberman BA (2004). Finding communities in linear time: A physics approach. *European Physical Journal B*, 38 (2), p. 331-338.

Zulueta MA; Bordons M (1999). A global approach to the study of teams in multidisciplinary research areas through bibliometric indicators. *Research Evaluation*, 8 (2), p. 111-118.



Appendix 1: Department of Electronic Technology. Specialties (2000-2004)



Appendix 2: Department of Computer Science. Specialties (1995-1999)

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)