Links in science : linking network and bibliometric analyses in the study of research performance

Calero Medina, C.M.

Author: Calero Medina, Clara
Title: Links in science : linking network and bibliometric analyses in the study of research performance
Date: 2012-12-11

8 Conclusions and future prospects

8.1 Key questions

This thesis originated from the need to identify groups of related nodes within collaboration and citation networks. In the study of collaboration networks the main goal is to identify existing research groups, potential research groups, or patterns of collaboration. The analysis of citations networks through specific measures and metrics, on the other hand, makes it possible to identify main lines of research through the years. Thus, such analyses improve our understanding of the growth and decline of fields, including phenomena such as paradigm shifts and emerging research themes. Network measures and metrics also allow for the identification of important nodes (e.g., journals, articles) embedded in the citation network. We addressed three main questions in this thesis:

Can we identify communities, existing research groups, and potential research groups?

Can we identify main lines of research through the years and the articles that linked them into a research tradition that can be considered the backbone of the field?

Can we identify important nodes that play a key role in the citation networks?

In the next sections we will discuss our findings concerning answers to these questions and the necessary future research.

8.2 Results

In **Chapter 2** we presented a method for identifying research groups and potential research partners in scientific fields. We combined a bibliometric science map based on a co-word network with the analysis of a co-publication network. A first and important result of the study is that we have identified *functional* rather than '*physical'* groups. Following Seglen and Aksness's (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 shortterm scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". The groups were defined over a six-year period, which means that the group members had not necessarily worked together. In addition, identification of the members via the combination of author name and affiliation address made it possible for the same person to belong to more than one group. This was the case, for instance, with a researcher who moved from one organization to another and changed his line of research and as a result belonged to two different groups in the period of analysis. A second significant outcome of the study is the possibility to identify potential research partners. Combination of output similarity relations with co-author relations offers a way to detect groups working in the same areas but not co-publishing. A third important result of our approach was that we were able to deal with the problem of homonymous and synonymous author names[1]. The combination of author and address data in a publication allowed us to handle the homonymous names, while the network analysis made it possible to deal with the second category. The combined data enabled us to assign author names to specific researchers more accurately**.**

In **Chapter 3** we presented the case study of how to use publications data to analyze the organizational structure of a large university hospital. Translational research in a university hospital is deeply embedded within daily work activities; it is not limited to a specific hierarchical or technical entity but widely distributed across the entire organization. Thus, proper management is very important in order to facilitate the research activities. In the past years we have observed considerable advances in the development of methods for finding communities within networks, with a large number of different techniques under development. This study shows how bibliometric analyses can benefit from these developments and complement them, since the case studies provided an insight into what the identified groups mean by, validating the results with the opinions of experts involved. The case study presented in Ch. 3 shows how the combination of bibliometric indicators and collaboration analysis can help research managers of large organizations and university hospitals in particular to understand the way the organization behaves, in order to create the strongest possible research clusters.

[1] Homonymous names are two or more persons with the same author name, while synonymous names are two or more different author names referring to the same person.

Chapter 4 describes the results of an empirical study in which we explored the analytical potential of corporate research articles as a source of empirical information for describing structural patterns within multinational enterprises (MNE) in the bio-pharmaceutical industry worldwide, and to produce quantitative data on those research cooperation relationships at the level of countries and major bio-pharmaceutical firms. Given the overwhelming significance of basic research in the bio-pharmaceutical industry and the large quantity of corporate research papers produced each year, we believe that these publications reflect key characteristics of research cooperation patterns within the industry. The outcome revealed interesting empirical information, not only with respect to the organizational features of corporate research partnerships within and between companies, but also on the geographical distribution of these partnerships. The company-level breakdown of these cooperation patterns also reveals a variety of intra- and extra-firm research linkages, from which three main types of corporate research networks can be derived in terms of the intra-firm distribution of research partnerships: (a) centralized networks, (b) decentralized networks, and (c) gateway networks.

Chapter 5 we described a broad study of bibliometric characteristics of largest 386 universities worldwide in terms of number of publications, and of a (partly overlapping) set of 529 European universities. Rather than presenting a ranking, the study presents a statistical analysis of ranking data, focusing on more *general* patterns. Several aspects were compared: US universities with European universities; countries with a strong concentration of academic research activities in a relatively small amount of universities, with nations showing a more even distribution of research over universities; a ranking of universities based on indicators calculated for all research fields combined, with one compiled for a single field (oncology); general with specialised universities; and rankings based on a single indicator with maps combining social network analysis and a series of indicators. The study highlights important factors that should be taken into account in the interpretation of rankings of research universities based on bibliometric indicators. Moreover, it illustrates policy-relevant research questions that may be addressed in secondary analyses of ranking data. In this way, the study was aimed at contributing to a public information system on research universities.

In **Chapter 6** we followed the 'intellectual track' of a specific research concept, *absorptive capacity* (AC), which had a high rate of diffusion through the fifteen years of analysis. With the bibliometric map further concepts (in terms of field-specific keywords) were found which are often related to theories and models associated with the main concept

(AC). Next, we used two other network-based citation-analysis techniques to find the main papers during these years, i.e., the articles that influenced the research for quite a time, and linked them to a research tradition that is the backbone of the 'Absorptive Capacity Field'. Our results show the potential of this methodology as a tool for unraveling the patterns hidden in a set of publications representing a field. The combination of bibliometric mapping with a detailed analysis of the citation network enables us to follow the influence of the introduction of a new concept in a specific research field.

Finally, in **Chapter 7** we presented a method for analyzing the 'citation environment' of a journal. Based on a bibliometric perspective of journal performance, our goal was to provide a fast but nevertheless comprehensive overview of the most important related journals for a given journal in terms of citations given and received. The method introduced in this chapter enabled us to establish the important journals in the citation environment of a given journal, their degree of importance, and the position they occupy in the network.

8.3 Answers to key questions

The answer to the first key question formulated in Section 8.1 is linked with the studies presented in Chapters 2 to 5. We have identified functional research groups embedded in a field (**Chapter 2**) and embedded in an organization (**Chapter 3**), together with potential research groups in a field (**Chapter 2**). We have found broader communities: groups of universities that collaborate based on geographical proximity (**Chapter 5**), and **(Chapter 4)** patterns of intra-firm and extra-firm collaboration.

Chapter 6 is linked to the second question, since in the study described there we identified a main line of research through the years and linked it to a research tradition that can be considered the backbone of the field.

Also in **Chapters 6** together with **Chapter 7,** we identified important nodes that play key roles in two types of citation networks. Thus, these chapters are related with the third question. In **Chapter 6** we identified papers while in **Chapter 7** we identified journals in the relevant citation networks.

8.4 Future Prospects

In general we can say that the future prospects of research as described in this thesis are strongly connected to the reinforcement of the applicability of quantitative studies of science and technology. This is particularly the case for our understanding of knowledge transfer in science and technology, and of directly related themes such as evaluation of research performance, knowledge diffusion, and growth of fields which may be the new sources for innovation. The study of these issues will benefit from the ongoing advances in measures and models of networked systems (citation-based and related networks in our case). They will contribute to a better understanding of the growth and decline of fields, including phenomena such as paradigm shift, emerging research themes, and the establishment of new institutions. Network analysis based on conceptual linkages will substantially improve the mapping of fields in science and technology, and the identification of emerging R&D themes and their actors.

We intend to keep working on the detailed structural properties of citation and collaboration networks. Many networks are characterized by hubs, i.e., nodes of high degree, see for instance Barabasi & Albert (1999); van Raan (2008). Highly cited publications evidently function as hubs, as they are the expression of the phenomenon of preferential attachment in citation networks (Jeong, Neda, & Barabasi (2003)). Mapping of interrelated entities makes it possible to study the topology of complex networks. In science such entities are publications, citations (van Raan, 2000), journals (Bergstrom, West, & Wiseman (2008)), institutes, and authors (Börner, Maru, & Goldstone, 2004). The search for hidden regularities and mathematical expressions to describe them is important because it may reveal the laws underlying the dynamics of complex networked systems (Leicht, Clarkson, Shedden, & Newman, 2007). Most complex networks are the results of a growth process (van Raan (2000), Newman (2001)). Science is an almost perfect example: a dynamical system that evolves through the addition and deletion of nodes and linkages, i.e., by new publications, their references, and newer publications citing older ones. Finding the dynamic rules that govern growth processes will lead to a better understanding of the resulting macroscopic, static properties of networks. To uncover the structure of network growth a rigorous mathematical model is needed. This may shed more light on problems such as the universality of networked systems, classification of networks, hierarchies, and the emergence of clusters, modules, and communities. The ensemble of modules represents highly interlinked communities (Rosvall & Bergstrom, 2007, 2008). How this modularity emerges is one of the basic questions in the study of network dynamics. Defining the relevant aggregation levels is important in order to understand the relation between citation networks and the impact of authors, and will enable us to find the life lines of science: what was a real breakthrough in the past? Interactions within and between clusters may change, for instance because of the development of a new, interdisciplinary field and its transformation into a mature and stand-alone discipline (Rosvall & Bergstrom, 2010). It is also important to define measurable quantities to describe interactions between time-dependent processes and static topology in the formation of complex networks. Understanding the regulatory and feedback mechanisms connecting various networks is one of the most ambitious goals in network research. Science offers an ideal target to tackle this problem because of the vast amount of data we have available and the presence of clearly observable quantities.

In line with the work described in this thesis we highlight especially the importance of mechanisms connecting citation- and co-authorship networks for the purpose of investigating the role of groups of researchers in the exchange and transfer of knowledge. As mentioned in the introduction of this thesis, the interconnections between scientific publications (e.g. citations given and received from one paper to another) and inside them (e.g. researchers co-authoring papers) allow us to study the way in which scientists create and share new knowledge. Citation networks of scientific publications can be viewed as composed of hierarchically layered networks. The lower network (basic network) consists of citations between scientific publications. A hierarchical step higher than the basic network is the network of citations between researchers. And a further step higher is the network of citations between research groups. Research groups form a crucial aggregation level because they represent the real work floor of science. The exchange of knowledge of research groups measured through the exchange of citations is part of our future interest.

The above issues are linked to another important problem: the identification and definition a research group. Given the large number of empirical studies conducted by CWTS, we have ample information about organizational structures of research institutions, so we can define a research group within the parent organization. In this case the nodes in the higher (third) network are defined by the organizational data. Thus, research groups form an aggregation of organizationally related publications, which is different from bibliometrically related (e.g., co-author based) publications. In other words, basic elements (nodes) of a lower network may also cluster in another network than their own organizational structure. As far as the bibliometrically related publications concerns, we can identify research groups by looking at authors in co-publication networks. Thus, the co-publication network itself represents a modular structure of co-author groups (Girvan & Newman, 2002; Newman 2004; Newman & Girvan 2004), not necessarily the same as the organization-based research groups to which the authors are affiliated. This bibliometric network and its modules are often more complex than the formal organization, particularly in interdisciplinary research.

Another important question is the translation of the commonly used bibliometric indicators into 'topological' properties of both the lower, basic network as well as of the higher-level networks. For instance, the number of citations of a group is the in-degree of the group in the higher network of groups, and the h-index is a specific variant of the total number of citations at a specific aggregate level (author, group). The impact of a research group in bibliometric terms is the ratio of the number of citations per publication of the group, and the number of citations per publication for the field(s) in which the group publishes. This field-normalized impact represents as it were the fitness of a group as a node in the higher network. The nominator can be seen as a field-specific property of the higher network that encompasses all research groups in science. However, recent studies have shown that normalized indicators are only mathematically consistent if this normalization is not on an aggregate level, but on the lower basic level. Further research is necessary to understand this in the context of network structures.
Furthermore, we intend to go a step further in our attempts to explain how *bibliometrically* related research groups emerge and evolve within the dynamic system of the entire scientific network. This issue is strongly related to cumulative advantage processes which can be analyzed together with other processes that sociologists have studied and found to be important (Powel et all., 2005). For instance, one of these processes is *homophily* (McPherson and Smith-Lovin, 1987), the process by which people who have common characteristics stick together. The study of this phenomenon has also been called 'assortative mixing in networks' (Newman, 2002; Newman, 2003; Newman and Park, 2003) in which the probability of two nodes being connected by an edge depends on specific similarity properties of the nodes. Another interesting process is what Powel et al (2005) have defined as *following the trend*: the network expansion follows a herd-like behaviour, either in response to external pressures, or through what they called 'imitative behaviour'. Finally, the above authors defined the process of *multiconectivity*, that reflects a preference for variety, for moving in different communities and interaction with heterogeneous partners suggesting a search for novelty.

Undoubtedly the research of complex networked systems will benefit from the vast amount of bibliometric data and from the characteristics of bibliometric constructs such as indicators and maps.

References

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

Bergstrom, C.T., West, J.D.and Wiseman, M.A. (2008). The Eigenfactor Metrics. Journal of Neuroscience, 28(45), 11433–11434.

Borner, K., Maru, J.T. and Goldstone, R.L (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 101 (Suppl 1), 5266.

Girvan, M. and Newman, M.E.J. (2002).Community structure in social and biological networks. Proceedings of the National Academic of Sciences of the USA 99, 7821-7826.

Jeong, H., Neda, Z. and Barabasi, A.L. (2003). Measuring preferential attachment in evolving networks. EPL (Europhysics Letters), 61(4):567–572.

Leicht, E.A, Clarkson, G., Shedden, K. and Newman, M.E.J. (2007). Large-scale structure of time evolving citation networks. The European Physical Journal B, 59(1), 75–83.

McPherson, M. and Smith-Lovin, L. (1987). Homophily in Voluntary Organizations. American Sociological Review 52, 370–79.

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. Physical Review E, 64(2), 25102.

Newman, M.E.J. (2002). Assortative mixing in networks. Physical Review Letters 89, article no. 208701.

Newman, M.E.J. (2003). Mixing patterns in networks. Physical Review E 67, article no. 026126.

Newman, M.E.J. (2004). Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, 2004.

Newman, M.E.J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* 69, article no. 026113.

Newman, M.E.J. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E* 68, article no. 036122.

Powel, W., White, D., Koput, K. and Owen-Smith J. (2005). Network dynamics and field Evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110, 1132-1205.

van Raan, A.F.J. (2000). On growth, ageing and fractal differentiation of science. Scientometrics, 47(2), 347–362.

van Raan, A.F.J. (2005). Reference-based publication networks with episodic memories. Scientometrics, 63(3), 549–566.

van Raan, A.F.J. (2006). Performance-related differences of bibliometric statistical properties of research groups: cumulative advantages and hierarchically layered networks. Journal of the American Society for Information Science and Technology, 57(14), 1919–1935.

van Raan, A.F.J. (2008). Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of research groups. Journal of the American Society for Information Science and Technology, 59(4):565–576, 2008.

Rosvall, M. and Bergstrom, C.T. (2007). An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences, 104(18), 7327–7331.

Rosvall, M. and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4), 1118–1123.

Rosvall, M. and Bergstrom, C.T. (2010). Quasi-correspondence analysis on scientometric transaction matrices. PLoS ONE, 5(1):e8694.

Seglen, Per O. and Aksnes Dag W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, Scientometrics 49 (1), 125-143.