Scientometrics (2010) 85:377–386
DOI 10.1007/s11192-010-0218-4

Shaping the European research collaboration in the 6th Framework Programme health thematic area through network analysis

José Luis Ortega · Isidro F. Aguillo

Received: 30 November 2009 / Published online: 1 April 2010

Abstract This paper aims to analyse the collaboration network of the 6th Framework Programme of the EU, specifically the “Life sciences, genomics and biotechnology for health” thematic area. A collaboration network of 2,132 participant organizations was built and several variables were added to improve the visualization such as type of organization and nationality. Several statistical tests and structural indicators were used to uncover the main characteristic of this collaboration network. Results show that the network is constituted by a dense core of government research organizations and universities which act as large hubs that attract new partners to the network, mainly companies and non-profit organizations.

Keywords Scientometrics · 6th Framework programme · Research collaboration · Network analysis

Introduction

The progress made in data analysis and computing has allowed to study in depth the structural relationships in complex environments such as the Web (Barabasi and Albert 1999), disease spreading (Pastor-Satorras and Vespignani 2001) and trophic dynamics (Polis and Strong 1996). However, the scientific activity could be also described as a complex system in which several agents (industry, university, government, etc.) interact in an environment subject to multiple variables. The use of structural analyses in R&D has made possible to understand collaboration phenomena in scientific journals (Newmann 2001;

J. L. Ortega (✉)
R&D Analysis, Vice-Presidency for Science and Technology, CSIC, Serrano, 113, 28002 Madrid, Spain
e-mail: jortega@orgc.csic.es

I. F. Aguillo
Cybermetrics Lab, CCHS-CSIC, Albasanz, 26-28, 28037 Madrid, Spain
e-mail: isidro.aguillo@cchs.csic.es

Barabasi et al. 2002; Wagner and Leydesdorff 2005), citation network among papers (Small 1999) and journals (Leydesdorff 2004) or relationships between patents (Valverde et al. 2007).

The European R&D system is strongly supported by the EU Framework Programmes, as relevant instruments for the building and strengthening of the European Research Area (ERA). These programmes assume the collaborative research as a principal feature of the ERA system in which the projects must to be carried out by several organizations from different countries and sectors. This networking environment provides a great opportunity to understand how these relationships are built, what are the main actors and their role and how the network operates in order to improve the EU R&D system.

Previous works have already analysed the collaborative networks of the European programmes. Breschi and Cusmano (2004) studied the R&D joint ventures of the 3rd and 4th Framework Programmes. They found that there is a preferential attachment phenomenon (Barabasi and Albert 1999) between both calls. Barber et al. (2006) studied from the second to the fifth framework programmes, confirming that these networks have scale-free properties such as power law degree distributions, small diameters and high clustering. Roediger-Schulga and Dachs (2006) found significant differences in two EU sub-programmes. They detected that while the telecommunication programme had more industrial partners and require greater funding; the agricultural one was dominated by public research institutions and attract less income. These differences between research programs was analysed by Cabo (1999) as well. Roediger-Schulga and Barber (2007), using the same data set, visualized the first five EU Framework programmes, showing that the backbone of the network is shaped by large scientific organizations.

Objectives

This paper aims to explore how the different research partners are related among them in the 6th Framework Programme of the EU, specifically, the "Life sciences, genomics and biotechnology for health" thematic area. We try to know what the role of each participant is in these programmes and how the firms, universities, governments and non-profit organizations interact between them. Observing those structural relationships, we also intend to estimate, using multiple regression model, if these structural indicators might explain and in what extent the percentage of funding that an organization receives.

Methodologically, we attempt to explain how the collaborative research processes are carried out in a research programme. The use of social network analysis (SNA) tools could be a suitable means to understand how the partner relationships are established when an EU project is executed.

Methods

Data

We have modelled a one-mode network set up by 2,132 organizations. They participate in 601 research projects belonging to the "Life sciences, genomics and biotechnology for health" thematic area from the 6th Framework Programme of the EU. These data were obtained through the Centre for the Development of Industrial Technology (CDTI), the Spanish public body depending of the Ministry of Science and Innovation in charge of promoting and funding innovation and technological development. CDTI supplied us an own database from eCORDA data (eCORDA, 2010). The above *ad-hoc* database contains the list of organizations (name, nationality and type of organization) which participate in each project (code, total cost and task) of the health thematic area. Participation table includes subvention, percentage of subvention, percentage of participation and role. However, a confidentiality rule only lets us to operate with aggregated data and percentages.

A normalization process was carried out to adapt the name of each institution to a standard name in English, removing different variations of the same name in different languages. We also removed acronyms, except when they are better known than their extended name, i.e., INSERM, IRCCS, etc. This normalization process reduced to 17% the number of organizations. Unfortunately, the project grant agreements are not always signed by the direct responsible (e.g., a laboratory, research group, etc.) of the project but by the main responsible of their research institution (e.g., the president of a research council, the rector of a university, etc.). This does not allow us to study these research relationships at the level of institutes or laboratories. Therefore, large institutions such as CNRS, CSIC or CNR are studied as one.

Several variables were included in order to add information about the network configuration and to design different analysis and relationships between variables and institutions. Nodes size shows the percentage (%) of funds allocated to each organization while the arc width indicates the number of projects in common with other organization. Each colour represents the country of each organization. The shape of the nodes shows the type of organization according to the institutional classification of the Frascati Manual (OECD 2003), being:

- Governments: All departments, offices and other bodies which furnish those common services which cannot otherwise be conveniently and economically provided, as well as those that administer the state and the economic and social policy of the community. It includes NPO controlled and mainly financed by government. They are represented by a triangle in the graph.
- Universities: All universities, colleges of technology and other institutions of post-secondary education, whatever their source of finance or legal status. It also includes all research institutes, experimental stations and clinics operating under the direct control of or administered by higher education institutions. A circle is used to show the universities in the map.
- Firms: It includes all firms, organisations and institutions whose primary activity is the market production of goods or services for sale to the general public, and the private non-profit institutions which mainly works for them such as trade associations, chambers of commerce, or those who are mainly funded (more than the 50%) by their commercial activity. For example, the Pasteur Institute is classified in this category because although it is a non-profit organization it obtains their income mainly through selling vaccines. Firms are described in the graph as squares.
- Non-Profit Organizations (NPO): private non-profit institutions not included in the above categories, as well as private individuals. We used a diamond to show the NPO in the map.

Network analysis

The software program Pajek 1.02 (Nooy et al. 2005) was used to build and visualise the network, while the Fruchtermann–Reingold algorithm (1991) was used to energize it.

Several network indicators and measurements were extracted from the network using Ucinet 6 (Borgatti et al. 2002). The following indicators were used in this study:

- Degree Centrality (k): It measures the number of lines incident to a node (Freeman 1979). This can be normalized (nDegree) by the total number of nodes in the network. Degree Centrality allows detecting organizations that have a high collaboration degree with different institutions, showing a high activity in the research programmes.
- Freeman's Betweenness centrality (C_B): the capacity of one node to help to connect those nodes that are not directly connected between them (Freeman 1980). Its normalization is the percentage over the total number of nodes in the network. From a scientometric point of view, this measurement enables us to detect hubs or gateways that connect different organizations to the core of the networks, showing the capability of certain institutions for attracting partners to the research programmes.
- K-Core: a sub-network in which each node has at least degree k. K-Cores allows us to detect groups with a strong link density. In scale-free networks the core with the highest degree is the central core of the network, detecting the set of nodes where the network rests on (Seidman 1983).

Some statistical tests were also used to contrast differences between types of organization according to their roles on the network.

- Kruskall–Wallis H test (1952) detects if n data groups belong or not to the same population. This statistic is a non-parametric test, suitable to non-normal distributions such as the power law distributions observed in scientometrics distributions.
- Dunn's post test (1961) compares the difference in the sum of ranks between two columns with the expected average difference (based on the number of groups and their size). It is used after the Kruskall–Wallis or Friedman test. The Dunn's test shows which samples are different.

Regression analysis

Several regression models were carried out in order to estimate and quantify the relationship between programme variables (funding) and network indicators (betweenness centrality, degree centrality). Linear regression permits us to know if exists a relationship of dependence between variables and what the weight of each variable is in the model. Regression goes beyond correlation by adding prediction capabilities. Due to this we have decided to use a regression model better than a correlation in order to know which variables could estimate the funding.

Two assumptions on this model are necessary: the independence of the observations and the normality of the distribution. The first one states that none of the observations determine the following one. The second assumption obliges the variables to have a normal distribution which density function has to be symmetric. Due to this, the used variables in this study have been transformed to logarithm.

It is usual to detect multicollinearity between the predictor variables in multiple regression models, because they are highly correlated between them. This statistical phenomenon can be observed with some statistics. Tolerance is $1 - R^2$ for the regression of that independent variable on all the other ones, so the greater tolerance coefficients, the more independent the variables are. A score less than 0.2 indicates collinearity. The Variance-inflation factor (VIF) is the reciprocal of Tolerance and values more than 4 indicate collinearity.

These statistical tests were performed with SPSS 17 and XLStat 2008 statistical packages.

Results

Descriptive analysis

A previous descriptive analysis was performed to observe the most relevant organizations according to several indicators.

Table 1 shows the first ten organizations by the percentage of funding that they receive in this thematic area. The percentage of funding can be considered as a quality indicator that measures the strength and importance of their presence in the research programme. The top organizations are the Institute National de la Santé et de la Recherche Médicale (INSERM) with a funding ratio of 2.51%, followed by the Max Planck Society (2.39%) and the Karolinska Institute (1.99%). According to the number of projects, the most active organization is also the INSERM with 164 projects (27.3%), followed by the Centre National de la Recherche Scientifique (CNRS) (130 projects; 21.6%) and the Karolinska Institute (110 projects; 18.3%). These same organizations are the most collaborative as well, because they maintain the same position if we observe the number of their partners. However, if we normalized the percentage of funding with respect to the number of projects (Funding %/Projects), which will give us an idea of the weight of these organizations in each project, the Max Planck Society is the organization with the best ratio (0.026), followed by the Medical Research Council (0.02) and the European Molecular Biology Laboratory (0.019). It is also interesting to note the strong presence of French and British organizations in the first positions, which is comparable to previous results (Gusmao 2000; Roediger-Schulga and Barber 2007).

Network analysis

The partners network (Fig. 1) shows small-world network properties as its clustering coefficient ($C = 0.849$). It is considerable higher than the expected for a random network ($C = 0.0002$) (Watts and Strogatz 1998). Furthermore, its average path length ($l = 2.36$) is

Table 1 The 10 first organizations by their percentage of funding

Rank	Country	Organization	Funding %	Partners	Projects
1	France	INSERM	2.51	956	164
2	Germany	Max Planck Society	2.39	546	92
3	Sweden	Karolinska Institute	1.99	802	110
4	France	CNRS	1.83	741	130
5	International	European Molecular Biology Laboratory	1.53	444	78
6	United Kingdom	University of Oxford	1.53	642	81
7	United Kingdom	Medical Research Council	1.38	518	69
8	France	Institute Pasteur	1.15	474	65
9	The Netherlands	Leiden University	1.05	597	63
10	United Kingdom	Imperial College London	.96	641	59

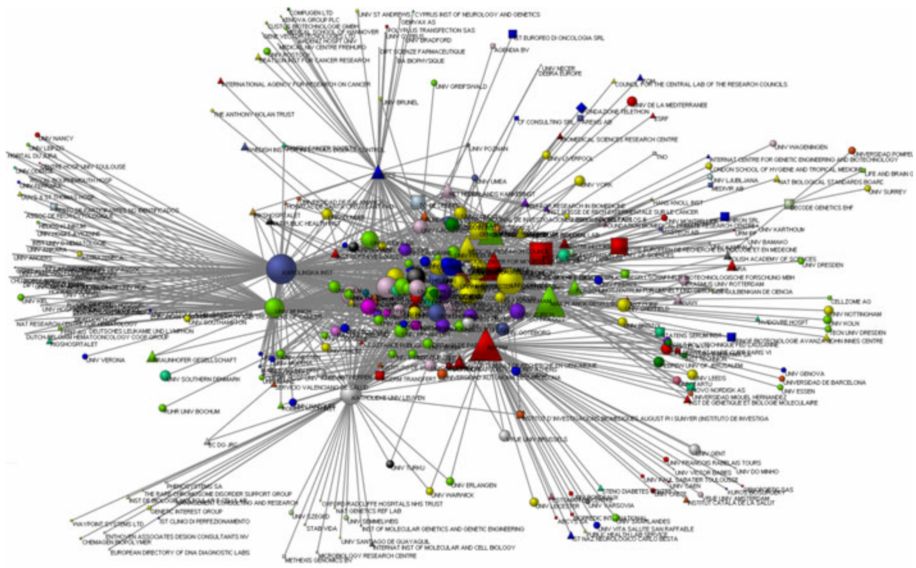

Fig. 1 Network of participant organization in the health thematic area ($n = 316$, arcs ≥ 5 partnerships)

also rather low. Visually, small-world network properties can be seen through the traversal links that run across the network, connecting distant clusters. The degree frequency distributions follow a power law trend ($\gamma = 1.64$) which enables us to state that this network owns scale-free properties as well (Barabasi et al. 2000).

Figure 1 shows the network of 316 organizations that have 5 or more partnerships with the same organization in the health programme. We have reduced the number of organizations in order to improve the visualization of the graphic and to see the principal characteristics of the network. Nevertheless, we have not been able to establish a larger cut-off because a network with more than ten projects in common will remove the 95% of the nodes. The *backbone* of the health thematic area shows a central core formed by Government institutions and Universities. This highly connected core ($k = 19$) of 28 organizations was detected using the k-Core technique. Germany (21%) and the UK (21%) contribute with the largest number of organizations to this core. However, France contributes only with a 10%, even though it is the country with most participants in the whole the health thematic area. This dense group is basically set up by universities (67%) and public research institutions (25%), while only one company is included in the core (Pasteur Institute). This results is confirmed by the Kruskall–Wallis H test. It detected significant differences (p-value < 0.0001) in the average of partners of each type of organization. Table 2 shows that the Government organizations have almost four times (86.47) more partners than the Company ones (23.31) and is far from the University category (58.55). The Dunn's post test shows that there are not differences between NPO and Company.

Figure 1 allows us to observe the presence of large hubs that attract organizations to the research programme. These hubs were identified with the Freeman's betweenness centrality and ranked in Table 3. The principal hubs that connect organizations to the main core of the research programme are government's research institutions and universities, some of which are central organizations in the research system of their respective

Table 2 Sample grouping according to the number of partners (Dunn's post test)

Sample	Frequency	Mean	Standard deviation	Groups		
Company	140	23.31	27.03	A		
NPO	1,028	28.91	41.82	A		
Government	406	86.47	115.43		B	
University	548	58.55	94.66			C

Table 3 The 10 first organizations by their centrality betweenness

Rank	Country	Organization	Betweenness
1	France	INSERM	0.079
2	Sweden	Karolinska Institute	0.052
3	France	CNRS	0.050
4	United Kingdom	University of Oxford	0.036
5	United Kingdom	Imperial College London	0.032
6	Germany	University of Munich	0.029
7	The Netherlands	Leiden University	0.028
8	Sweden	Lund University	0.025
9	Germany	Max Planck Society	0.025
10	Czech Republic	Academy of Sciences of the Czech Republic	0.024

countries. INSERM and CNRS in France, the Max Planck Society in Germany or the Academy of Sciences of the Czech Republic are examples of highlighted hubs in the network. Using the Kruskall–Wallis H test, significant differences (p-value < 0.0001) were detected in the mean betweenness centrality of the different types of organization (Table 4). Thus, the Government ($C_B = 0.002$) and University organizations ($C_B = 0.001$) have the highest mean betweenness centrality in the network, while the NPOs and Companies ($C_B = 0.0$) do not have any mediator property.

Regression analysis

A regression analysis was done to know which and in what manner these structural variables (degree centrality and betweenness centrality) would affect to the amount of funding that each organization gains in this research thematic area. Three variables were used in the model: total number of partners, number of coordinated projects and total number of projects. Betweenness centrality was rejected by the model because this variable showed strong collinearity with the variable total number of partners (degree centrality). The rest of variables showed acceptable levels of Tolerance and VIF. Then we can accept the absence of collinearity in the model.

Table 5 shows the obtained results form the regression model with an explanation of 50%. It means that the funding of an organization in the 6PM is determined for 50% by the number of projects, partners and coordinated projects. Multiple regression model assess in which proportion these variables explain the funding and what is the contribution of each one of them?

Table 4 Sample grouping according to the centrality betwenness (Dunn's post test)

Sample	Frequency	Mean	Standard deviation	Groups		
NPO	140	0.000	0.001	A		
Company	1,028	0.000	0.001	A		
Government	406	0.002	0.005		B	
University	548	0.001	0.005			C

Table 5 Multiple regression analysis model of the percentage of funding (Adjusted $R^2 = 0.50$)

Model	Unstandardized coefficients		Standardized coefficients	t	Sig.	Collinearity statistics	
	B	Std. error	Beta			Tolerance	VIF
(Constant)	12.697	0.381		33.316	0.000		
Projects	1.035	0.213	0.453	4.862	0.000	0.292	3.421
Coordinators	1.623	0.242	0.483	6.714	0.000	0.490	2.040
Partners	−0.424	0.140	−0.238	−3.023	0.003	0.409	2.443

The obtained model interprets that if the number of projects increases 10%—maintaining constant the other variables—could cause upward of 10.3% the incomes, while a similar increase in coordinated projects would go up 16% the funding. However, if the number of partners increases then the funding would go down. It seems contradictory because a simple regression model shows that the increase in partners could raise the funding (6.7%). This is due to a strong correlation between partners and projects ($R^2 = 0.59$), which affects to the estimation of the funding in the multiple regression model. Therefore, results suggest that the increase in partners is positive only if the number of projects increases as well. So, the rise of partners favours the participation in new projects, and therefore helps to get more funding.

Discussion and conclusions

The analysis of the participation in the health thematic area of the 6FP has allowed to describe the principal actors in this research programme according to several indicators. INSERM, the Max Planck Society, the CNRS and the Karolinska Institute are the most highlighted organizations in the thematic area, because they obtain the largest proportion of funding and participate in the largest number of projects. These organizations are, in the great majority of the cases, central institutions in the biomedicine research system of their countries. Thus, INSERM and the CNRS are the main research centres in France, Max Planck Society in Germany and the Karolinska Institute in Sweden. The K-Core allowed us to identify the nucleus of the network, which is mainly set up by universities and government organizations such as research councils and public research bodies. These organizations are the most qualified partners to develop a research project, because they have gained extensive experience and knowledge participating in previous research projects. However, studies on different thematic areas (Cabo 1999; Roediger-Schluga and Dachs 2006) have shown that the core changes according to the research field. Thus in technical areas there is a higher proportion of large companies in the core, maybe because these sectors are more interested in development-related projects with a strong business profit orientation. Nevertheless, the health area is supported by basic research, which is developed by universities and research centres probably due to the social relevance of the health. This causes that the most interested agent in performing health research are public bodies (McMillan et al. 2000).

The observed role of companies and NPOs in the health thematic area is rather peripheral. Although Company is the largest set of organizations (48.44%), it is almost not found in the core of the network and it has the lowest partners and betweenness centrality mean. This lets us to suppose that the companies participate in few and specific projects oriented to their business line, and look for the support of universities and research centres, which are located in the core of the programme, to develop those projects. This also may be due to the fact that most of these companies are small-size bio companies—10% of the participants are small and medium enterprises (European Comission 2008)—born from the university (*spin-offs*) (George et al. 2002) and with an intensive activity in specialized areas (Biotechnology Industry Organization 2008). Maybe these particular characteristics of the biotechnology companies explain their peripheral situation.

The obtained results lead us to speculate an explanation of how the collaboration network of the health thematic area of the 6FP is formed. The network is constituted by a dense core of government research organizations and universities. They are the most outstanding research actors in the system. This causes a cumulative process (Price 1976; Barabasi and Albert 1999) in which these principal entities participate in more and more European research projects, gaining technical and knowledge resources. Most of them act as large hubs that attract new partners to the network. Their prestige and experience make possible that these new partners contact with them in order to develop a research project. Most of these new members are companies and NPO. They have a low participation degree because they are small size companies which are centred in a specific business line and they are focused in specific projects with defined partners.

The multiple regression model showed that it is more profitable to be a coordinator (16%) than not (10%), because the income is larger if an institution participates with that role. On the opposite, the number of partners causes the contrary effect reducing to 4% of income. This opposite result is because the more partners we have the less money we share, being the number of projects constant. Hence, the partners affect both the income and the projects. This strong relationship between projects and partners suggests that the increase of partners would raise the likelihood to participate in new projects and therefore to obtain more funding from the programme. So we may conclude that to contact with new partners helps to improve the income as long as that involves more projects.

Acknowledgments We wish to thank the R&D Framework Programmes Department of the Centre for the Development of Industrial Technology (CDTI) of Spain for their support and the supply of 6th EU Framework Programme data.

References

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

Barabasi, A. L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A*, *281*(1–4), 69–77.

Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*(3–4), 590–614.

Barber, M. J., Krueger, A., Krueger, T., & Roediger-Schulga, T. (2006). The network of European research and development projects. *Physical Review E Statistical Nonlinear and Soft Matter Physics, 73*(3), 1–13.
Biotechnology Industry Organization. (2008). *Technology, talent and capital: State biosciences initiatives 200*8. Washington: Battelle. Retrieved October 07, 2009 http://bio.org/local/battelle2008/State_Bioscience_Initiatives_2008.pdf.
Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.
Breschi, S., & Cusmano, L. (2004). Unveiling the texture of a European research area: Emergence of oligarchic networks under the EU framework programmes. *International Journal of Technology Management, 27*(8), 747–772.
Cabo, P. G. (1999). Industrial participation and knowledge transfer in joint R&D projects. *International Journal of Technology Management, 18*(3–4), 188–206.
Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(1961), 54–64.
European Commission. (2008). FP6 final review: Subscription, implementation, participation. Brussels: Research Directorate General. http://ec.europa.eu/research/reports/2008/pdf/fp6-final-review.pdf.
Freeman, L. C. (1979). Centrality in networks: I. conceptual clarification. *Social Networks, 1*, 215–239.
Freeman, L. C. (1980). The gatekeeper, pair-dependency, and structural centrality. *Quality and Quantity, 14*, 585–592.
Fruchtermann, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software Practice and Experience, 21*(11), 1129–1164.
George, G., Zahra, S. A., & Robley Wood, D. (2002). The effects of business-university alliances on innovative output and financial performance: A study of publicly traded biotechnology companies. *Journal of Business Venturing, 17*(6), 577–609.
Gusmao, R. (2000). Developing and using indicators of multilateral S&T cooperation for policy making: The experience from European research programmes. *Scientometrics*, *47*(3), 493–514.
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621.
Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in Journal Citation Reports. *Journal of Documentation, 60*(4), 371–427.
Mcmillan, G. S., Narin, F., & Deeds, D. L. (2000). An analysis of the critical role of public science in innovation: The case of biotechnology. *Research Policy, 29*(1), 1–8.
Newmann, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, 64*(1), 016131.
Nooy, W., de Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge, UK: Cambridge University Press.
OECD. (2003). *Frascati manual 2002*. Paris: OECD Publishing.
Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters, 86*(14), 3200–3203.
Polis, G. A., & Strong, D. R. (1996). Food web complexity and community dynamics. *The American Naturalist, 147*(5), 813–846.
Price, D. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*, 292–306.
Roediger-Schulga, T., & Barber, M. J. (2007). *R&D collaboration networks in the European framework programmes: Data processing, network construction and selected results* (pp. 1–39). Maastricht: United Nation University.
Roediger-Schulga, T., & Dachs, B. (2006). *Does technology affect network structure? A quantitative analysis of collaborative research projects in two specific EU programmes* (pp. 1–29). Maastricht: United Nation University.
Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*, 269–287.
Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.
Valverde, S., Sole, R. V., Bedau, M. A., & Packard, N. (2007). Topology and evolution of technology innovation networks. *Physical Review E, 76*(5), 056118.
Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy, 34*(10), 1608–1618.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, *393*, 440–442.