Journal of Informetrics 9 (2015) 809–825

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

How to become an important player in scientific collaboration networks?

Ashkan Ebadi[a,*], Andrea Schiffauerova[a,b]

[a] Concordia Institute for Information Systems Engineering (CIISE), Concordia University, 1515 Ste-Catherine Street West, Montreal, QC, Canada H3G 2W1

[b] Department of Engineering Systems and Management, Masdar Institute of Science and Technology, Masdar City, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form 30 July 2015
Accepted 4 August 2015
Available online 3 September 2015

Keywords:
Network structure
Collaboration
Statistical analysis
NSERC
Canada

ABSTRACT

Scientific collaboration is one of the important drivers of research progress that supports researchers in the generation of novel ideas. Collaboration networks and their impact on scientific activities thus already attracted some attention in the research community, but no work so far studied possible factors which can influence the network positions of the researchers at the individual level. The objective of this paper is to investigate various characteristics and roles of the researchers occupying important positions in the collaboration network. For this purpose, we focus on the collaboration network among Canadian researchers during the period of 1996 to 2010 and employ multiple regression models to estimate the impact on network structure variables. Results highlight the crucial role of past productivity of the researchers along with their available funding in determining and improving their position in the co-authorship network. It is shown that researchers who have great influence on their local community do not necessarily publish high quality works. We also find that highly productive researchers not only have more important connections but also play a critical role in connecting other researchers. Moreover, although mid-career scientists tend to collaborate more in knit groups and on average have higher influence on their local community, our results specifically highlight the important role of young researchers who occupy mediatory positions in the network which enable them to connect different communities and fuel information transmission through the network.

1. Introduction

Recent progress in information technologies has cut the world-wide distances enabling researchers to get in contact easier. Hence, nowadays no specific border can be defined for scientific activities in a way that researchers have formed a global community aiming to advance the level of knowledge. Concurrently, the nature of the science has become more complex and inter-disciplinary which encourages scientists to be more collaborative in order to increase their scientific productivity, to get access to new knowledge and financial resources, *etc.* Katz and Martin (1997) define scientific collaboration as the process through which researchers with a common goal work together to produce new scientific knowledge. The importance of collaborative research is now acknowledged in scientific communities (Brad Wray, 2006). Through collaboration researchers get access to an often informal network of scientists that may facilitate knowledge and skill diffusion (Tijssen, van Leeuwen,

* Corresponding author. Tel.: +1 352 745 4468.
E-mail address: a_ebad@ciise.concordia.ca (A. Ebadi).

http://dx.doi.org/10.1016/j.joi.2015.08.002
1751-1577/

& Korevaar, 1996; Tijssen, 2004). Although it is not easy to quantify scientific collaboration, co-authorship has become the standard way of measuring collaboration since it is considered as a better sign of mutual scientific activity (De Solla Price, 1963; Ubfal & Maffioli, 2011). Co-authorship networks, as one of the main forms of scientific collaboration (Abbasi, Altmann, & Hwang, 2010), evolve over time. This evolution might reflect the growth/decay of a research subject, community or even a scientific field (Huang, Zhuang, Li, & Giles, 2008). This evolution and changes can be also seen at the nodes level (*i.e.* researchers in the co-authorship networks) where researchers' positions and their importance within their community and/or the whole collaboration network might also change over time. Position of a node in a network depends both on its direct and indirect connections with the other nodes (Mattsson & Johanson, 1992).

Due to the growing large number of researchers and their co-authorship links, scientific collaboration networks are among the complex ones (Abbasi, Hossain, & Leydesdorff, 2012). Role of a researcher (node) in a network can bring some advantages to the researcher (*e.g.* better access to knowledge sources, political factors, awareness of potential projects, *etc.*), and the surrounding community. This becomes more interesting as one notes that the roles of nodes in a network might change over time (Abbasi *et al.*, 2012). Barabási and Albert (1999) showed that a new node in a network will be linked to the other nodes with large number of connections (higher degree centrality) with a higher probability. This indicates the importance of the highly connected nodes in a network. This is also confirmed by Moody (2004) who showed that authors who are new in a scientific network are more likely to get connected to highly reputable authors with many collaborators thus making the surrounding community of the reputable researcher denser. On the other hand, there exist studies indicating that getting connected to high performing nodes (researchers, organizations, *etc.*) can affect the performance of the connecting node. For example, Mote (2005) analyzed the impact of inter-organizational complexity on the research output of 20 projects in national labs and found that groups that were connected to prolific organizations also showed higher performance. All of this highlights the importance of structural collaboration network positions in scientific and technological activities. Thus this paper specifically focuses on researchers' roles in their collaboration networks and assesses the impact of influencing factors.

The remainder of the paper proceeds as follows: Section 2 discusses the gaps in the literature and objectives of the research; Section 3 presents the data, methodology and the models; Section 4 presents the empirical results and interpretations; Section 5 concludes; and Section 6 discusses the limitations.

2. Research motivation and objectives

Scientific collaboration is more and more attracting the attention of researchers as the science is evolving toward a more complex and highly inter-disciplinary nature. The continuous growing trend of collaboration in terms of the number of co-authored papers has been widely confirmed in bibliometric studies (*e.g.* Grossman, 2002; Cronin, 2005). In addition, it has been studied in a vast number of different disciplines such as computer science, sociology, research policy, and philosophy (Sonnenwald, 2007), focusing on different aspects of collaboration. In a quite different study, Jiang (2008) presented an algorithm for detecting active researchers in scientific communities which is based on an abstract definition of collaboration cost and number of interactions between researchers. Their assumption of considering active researchers to be more attractive for collaboration partially confirms the importance of collaboration in scientific communities. Abbasi *et al.* (2010) used the three measures of researchers' collaboration network structure, number of collaborations and productivity of co-authors to quantify the collaboration activities of researchers. They proposed two indices, namely researchers collaboration (RC-index) and community collaboration (CC-index), which can be also used for detecting the best partners for a research project.

It is argued that the structure of the network can affect the collaboration patterns and scientific output (Ebadi & Schiffauerova, 2015a). Several studies assessed the impact of collaboration patterns and network positions on scientific activities and performance of researchers (*e.g.* Eslami, Ebadi, & Schiffauerova, 2013; Beaudry & Allaoui, 2012; Abbasi, Altmann, & Hossain, 2011) as well as their level of funding (*e.g.* Ebadi & Schiffauerova, 2015b) and found a positive relation in most of the cases. For example, Abbasi *et al.* (2011) focused on the impact of four network indicators (*i.e.* degree centrality, closeness centrality, betweenness centrality and eigenvector centrality) along with some other factors on the citation-based performance of researchers who were active in information systems field and found a positive relation between eigenvector and degree centralities and the performance of the target scholars. In another study, Abbasi *et al.* (2012) analyzed the impact of possessing various roles in co-authorship network in observing new researchers for collaboration. Their results suggest the higher importance of betweenness centrality as well as the degree centrality of an existing researcher in attracting new entrants. In addition, there are a number of studies that evaluated the impact of several influencing factors (*e.g.* funding, gender, scientific fields) on scientific collaboration and its patterns (*e.g.* Bozeman & Corley, 2004; Adams, Black, Clemmons, & Stephan, 2005; Gulbrandsen & Smeby, 2005; Rosenzweig *et al.*, 2008; Defazio, Lockett, & Wright, 2009). For more information, see the critical review of the literature by Ebadi and Schiffauerova (2013).

Although there are some studies that confirms the importance of structural network positions and relationships in business and scientific communities (*e.g.* Håkansson & Ford, 2002), to the best of our knowledge no study shows how one can possess such network positions by analyzing the impact of influencing factors on different network positions in scientific collaboration networks. In other words, network structure variables have been so far considered at the right hand-side of the equations, estimating their impact on various scientific activities or performance of the researchers, *etc.*

Apart from performance related factors and financial power, we hypothesize that career age and affiliation type of a researcher might help him/her to possess more influential network positions. We define an *influential researcher* as a highly

central and important researcher who can play a determinant role in his/her local or even global collaboration network. We will use the general term of *influential* in the rest of the paper as the definition might slightly differs for each of the network structure variables. Senior researchers might have on average a better established collaboration network and be more known within their own community as well as the surrounding ones. Therefore, we expect that senior researchers can possess some network positions easier than their young counterparts. Industrial researchers might have on average easier access to the financial resources as well as the required research equipment which can help them to not only speed up their research but to get in contact with other researchers easier. This makes industrial researchers as attractive nodes to connect to in collaboration networks, hence academic researchers are also eager to get connected to prolific industry teams/projects (Balconi & Laboranti, 2006). Therefore, industrial researchers might have a privilege in acquiring some influencing network positions.

Analyzing the impact of influencing factors of different types on researchers' network positions can not only highlight the determinant factors in playing different roles in the network but it can be also used as a guide for researchers who are willing to play such roles. Considering a scientific network as a set of researchers and their inter-relations, a change in a position of a researcher not only might affect his/her performance but the performance of researchers who are connected to him/her. Various network positions might provide a researcher with different privileges. For example, researchers who connect different clusters might gain an advantage in securing more research money as they have a control over the network and the flow of information. Another example would be researchers with high influence among their local community that might enable them to better conduct a research or to find an appropriate research partner easier.

This paper uses a large dataset of Canadian researchers active in natural sciences and engineering and conducts a statistical analysis to reveal the inter-relations among the selected influencing factors and various structural network positions at the individual level of the researchers. Social network analysis was used to calculate the centrality measures of the target researchers. In addition, several pre-processing stages were taken to assure the accuracy and quality of the data that will be explained later in the respective section. We expect different factors to be more important in possessing different network positions thus we inspect the determinant factors for each of the roles separately. In particular, our motivating research questions are: What are the most determinant factors in acquiring various positions in scientific collaboration networks? How the profiles of researchers with different network positions look like? Is it better to be affiliated with industry to play important roles in the network? Is possessing an important role in the collaboration network biased toward senior researchers?

The main contributions of the paper are: (1) focusing on network structure measures as dependent variables and estimating the impact of various influencing factors of different types at the individual level of researchers, (2) employing state-of-the-art techniques for cleaning and pre-processing the research data as well as using a unique procedure for retrieving articles which helped us to gather more accurate data, (3) using a (the most) comprehensive and clean database of Canadian researchers and performing the analysis at the country level of researchers involved in engineering and natural sciences.

3. Data and methodology

3.1. Data

The data for this research was gathered in three phases. We expected funding to be one of the important factors which affect the positions of researchers in the network. Since we were interested in the network positions of the Canadian researchers, the Natural Sciences and Engineering Research Council (NSERC) of Canada was selected as the source of funding data. The availability of the data as well as NSERC's role as the main federal funding organization in Canada, and the fact that almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC (Godin, 2003) were some of the reasons of such selection. In addition, NSERC lists full names of funded researchers that were helpful in our entity disambiguation procedure (explained later). Hence in the first phase, the funded researchers' data was extracted from NSERC. Several preprocessing modules coded in JAVA were applied on the collected funding data. First, the data was checked and corrected automatically for any special characters (*e.g.* French characters) and the names were parsed into first name and last name. Students were also removed from the data as the NSERC database originally contains both scholarships and grants. In the extracted data, funds were assigned to the principal investigator (PI) and all the other co-researchers were mentioned in the same record. As the next stage, NSERC funding that contained a principal investigator was equally divided among the PI and all the co-researchers. We validated the assumption of the equal division of the amount by holding interviews with experts who were randomly selected through a stratified sampling method. This resulted in 379,891 records of funded researchers. Since some researchers received several funds from NSERC in the same year, we created another complementary dataset through merging the records for a given researcher in a given year by adding up the amounts, making the set of (researcher, year) unique for each year. The final funding database contains 228,417 records of funded researchers within the period of 1996 to 2010. The data includes name of grantee, title of the research project, year, amount, researcher's affiliation, funding program, *etc*.

In the next phase, we used Elsevier's Scopus to collect all the information about the articles (*e.g.* co-authors, their affiliations, year of publication, annual citation counts) that were published by the funded researchers within the period of 1996 to 2010. We decided to focus on the period of 1996 to 2010 since the data coverage of Scopus (*e.g.* citation data) was better after 1996. To extract the articles, we did a full text search over the articles and fetched the ones that acknowledged

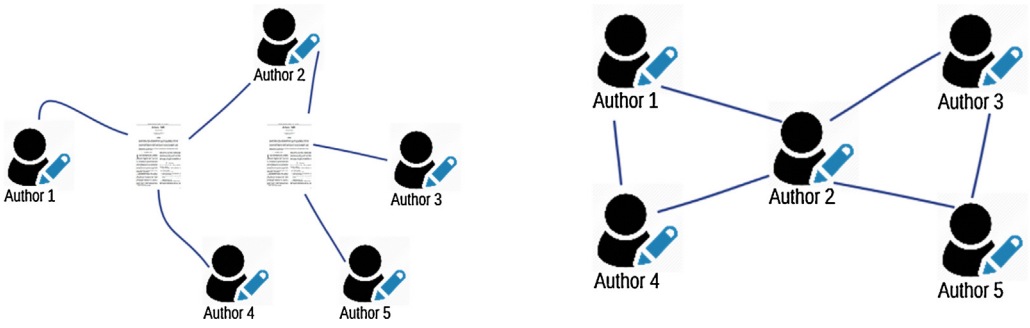

Fig. 1. (a) Two-mode co-authorship network, (b) One-mode co-authorship network.

NSERC support in the body of the articles. Using a list of keywords as the input (different formats of the way NSERC can be written and spelled), an automatic data collection procedure was designed and coded in JAVA for this purpose. This was a crucial step in gathering more accurate data since the common procedure in the similar studies is extracting the funded researchers' data and then gathering all the articles that were published by those researchers. Such procedure will most likely result in an over-estimation of the number of articles, as researchers usually use several sources of funding. The procedure that we followed was based on the assumption that all the NSERC grantees acknowledge the source of funding in the article. According to NSERC policies and regulations, researchers are required to acknowledge the source of funding in their publications. We held more than 30 interviews with randomly selected funded researchers and almost all of them approved such assumption and confirmed that they do acknowledge NSERC in their papers. Hence, 144,156 articles were extracted within the mentioned time interval. Several preprocessing steps were taken on the collected data, *e.g.* correcting special characters, parsing affiliations, and detecting the research area. Latent Dirichlet Allocation (LDA) technique, first introduced by Blei, Ng, and Jordan (2003), was used for keywords extraction from the title of the articles and research area detection. After performing the preprocessing stage, a crucial step was matching authors in the publications database with the funded researchers in the funding database. Two particular problems then arose: (1) To determine if for example "Alan Smith", "A. Smith" and "A. C. Smith" are all pointing to the same author, and (2) If "Alan Smith" in "Concordia University" is the same person as "Alan Smith" in "University of Toronto". For this purpose, machine learning entity disambiguation techniques were employed. NSERC data was almost clean containing the full names and affiliations of the funded researchers and Scopus data contained the current and past affiliations of authors. To perform the entity disambiguation between funding and publication databases, a similarity measure was defined based on name of researcher, his/her affiliations (including past affiliations) and research area. A JAVA program was coded implementing the Nearest Neighbor algorithm. The assignment procedure was designed semi-automatic due to the difficulties in entity disambiguation. That means the system asked user for a final decision on cases for which it could not obtain a high similarity score. At the end of this stage, the same Scopus-id was assigned to the matched records in funding and publication databases. The integrated database contains 174,773 records of disambiguated researchers within the period of 1996 to 2010. To have a proxy of the quality of the papers we used SCImago to collect the impact factor information of the journals in which the articles were published, *i.e.* SCImago Journal Rank (SJR). SCImago was chosen for two main reasons. First, it provides annual data of the journal rankings that enabled us to perform a more accurate analysis since we considered the impact factor of the journal in the year that an article was published, and not its impact in the current year. Second, SCImago is powered by Scopus that makes it more compatible with our articles database.

In the last phase of the data gathering procedure, we used Pajek software to construct the co-authorship networks of the target researchers for each of the single years of the examined time interval and to calculate the network structure variables at the individual level. For this purpose, a two-mode co-authorship network (De Nooy, Mrvar, & Batagelj, 2005) of researchers was first created in which nodes represent both authors and articles and articles are connected to their respective authors (Fig. 1a). Since authors were not directly connected to each other in the created two-mode network, as the next step we converted the created two-mode networks to one-mode networks in which two given authors are connected to each other if they have a joint article (Fig. 1b).

Five network structure variables were calculated on the created one-mode co-authorship networks, *i.e.* betweenness centrality, clustering coefficient, eigenvector centrality, closeness centrality and degree centrality. The first four variables were considered as dependent variables while the latter was regarded as one of the independent variables. The definiton of the network structure variables will follow in Section 3.2. The calculated network structure indicators along with the extracted journal quality measures were added to the final database, *i.e.* the target data. Fig. 2 shows the overall data gathering procedure. The models and variables that were used in this research are presented in the following section.

3.2. Analytical models

The main objective of this research is determining the influencing factors that help researchers to acquire various roles in scientific collaboration networks as well as their relation with the number of co-authors of researchers. Hence, we considered two general models including two different types of collaboration proxies as dependent variables. The first model considers

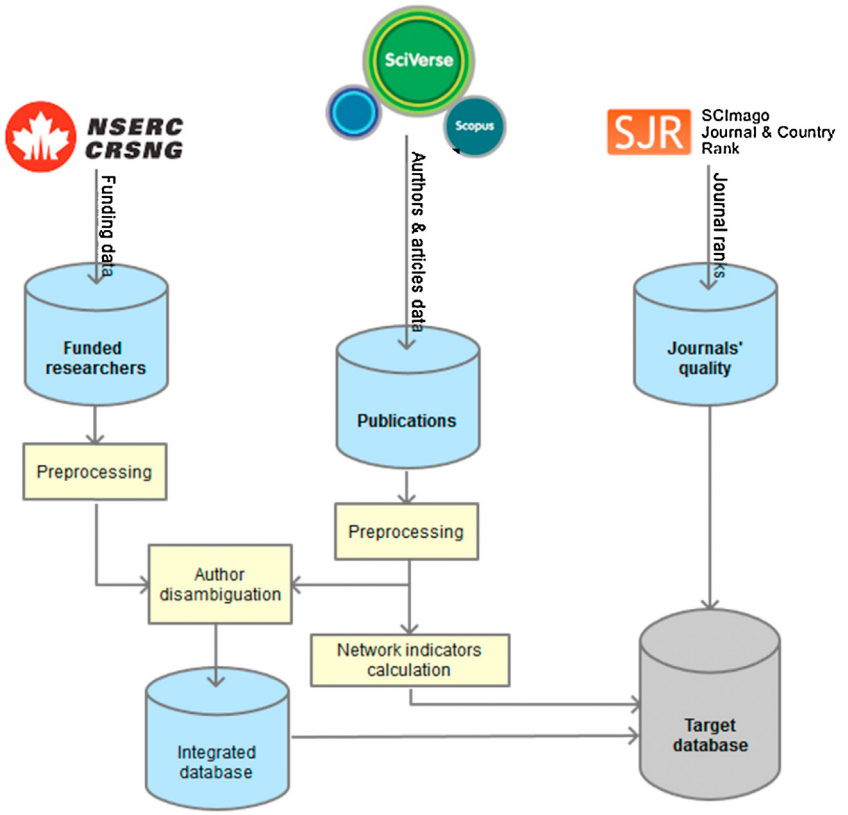

Fig. 2. The data gathering procedure in brief.

the average number of authors per paper as the collaboration measure (Model 1) while the second one (Model 2) considers various network structure indicators as dependent variables.

3.2.1. Model 1: The common collaboration proxy

In our first model, relationship between the selected influencing factors and the direct team size of researchers is investigated. The direct team size is calculated based on the number of authors per paper which is a common indicator of scientific collaboration and has been widely used in the literature (*e.g.* Beaver & Rosen, 1979; Rosenzweig *et al.*, 2008). We consider two slightly different dependent variables, one is based on the overall average number of co-authors for a given researcher (is referred to Model 1-a in the rest of the paper) and the other one focuses on the average distinct number of co-authors (Model 1-b). It should be noted that the distinct team size is different from the overall team size as well as the degree of a node. For example, suppose we have an author who has published articles jointly with three authors (not necessarily distinct authors), hence the degree of the author is three. Now, suppose the same author has published two articles, one of them in a co-authorship with author-A while the other one was written with authors A and B. Hence, the given author worked with three co-authors on two articles thus the overall team size is 1.5 (3/2). However, since he has worked with just two distinct co-authors (A and B) the distinct team size will be 1 (2/2). Comparing the results for Models 1-a and 1-b can reveal if the influencing factors are more correlated with the overall number of research partners or with the specific co-authors of given researchers. The model to be estimated in the reduced form is as follows (independent variables are explained in Section 3.3):

$$\begin{bmatrix} teamSize_i \\ disTeamSize_i \end{bmatrix} = \beta_1 \times avgFund3_{i-1} + \beta_2 \times noArt3_{i-1} + \beta_3 \times avgIf3_{i-1} + \beta_4 \times avgCit3_{i-1} + \beta_5 \times careerAge_i + \beta_6 \times dAcademia_i + \alpha_i \quad (1)$$

3.2.2. Model 2: Researchers' position in collaboration network

As discussed before, different network positions can bring various advantages to researchers. Model 2 specifically focuses on four important network measures (*i.e.* betweenness centrality (*bc*), clustering coefficient (*cc*), eigenvector centrality (*ec*),

and closeness centrality (*cl*)) and estimates the impact of the influencing factors in possessing and playing various network structural roles. We are interested in finding the set of the most important factors in playing each role in the collaboration network.

Betweenness centrality (*bc*) focuses on the role of intermediary individuals in a network. This measure identifies some important players, also called as *gatekeepers*, in the network who are able to bridge different communities. Such actors play an important role in knowledge and innovation diffusion as they have control over different clusters (sub-networks). Thus, being on the path of information pool and bridging different clusters might bring a strategic advantage to the gatekeepers in getting involved in new projects, finding partners or securing financial resources. These factors were our main motivations to study this role. Theoretically, betweenness centrality of node k is measured based on the share of times that a node i reaches a node j *via* the shortest path passing from node k (Borgatti, 2005). Hence, the more a node lies on the shortest path between any two other nodes in a network, the higher betweenness centrality it has which indicates the higher control that the node has over other two non-adjacent nodes (Wasserman, 1994). Hence, betweenness centrality of node k (bc_k) is defined as follows:

$$bc_k = \sum_{i \neq k \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \tag{2}$$

where σ_{ij} is the total number of shortest paths from node i to j and $\sigma_{ij}(k)$ is the number of shortest paths from node i to node j that contains node k.

Clustering coefficient (*cc*), also called *cliquishness*, counts the number of triangles in a given undirected graph to measure the level of clustering in the network. In other words, it is the likelihood that two neighbors of a node are also connected to each other (Hanneman & Riddle, 2011). Researchers with high clustering coefficient tend to cluster with other researchers resulting in tightly knit group collaboration with high number of connections among the team members. Thus, such researchers might benefit from the tight inter-connections in their groups to produce higher quality works by using the internal referring among the team members. Therefore, possessing this role can also bring some advantages to researchers which makes it interesting for our analysis. An intuition is to see this type of collaboration in some specific areas due to cultural issues or language factor. Watts and Strogatz (1998) define clustering coefficient based on a local clustering coefficient (*lcc*) for each node within a network. The definition of *lcc* is:

$$lcc_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered on node } i} \tag{3}$$

The denominator of the above formula counts the number of set of two edges that are connected to the node i (triples). The numerator counts the number of three nodes that are all connected to each other. The overall clustering coefficient is calculated by taking average of the local clustering coefficient of all the nodes within the network. Hence,

$$cc = \left[\sum_{i=1}^{n} {}^{LCC_i}\right] /_n \tag{4}$$

in which n denotes the number of vertices in the network. This measure returns a value between 0 and 1 in a way that it gets closer to 1 as the network interconnectivity increases (higher cliquishness).

Eigenvector centrality (*ec*) is based on the idea that the importance of a node in a network also depends on the importance of its connections. Hence, an actor has higher eigenvector centrality if it is connected with other important actors who are themselves occupying important positions in the network. In other words, eigenvector centrality measures how well connected an actor is in a network thus, it can be regarded as a measure of a researcher's influence in the network. Being connected to other highly important (central) researchers can bring a strategic and diplomatic power to a researcher which makes the role interesting for our analysis. Bonacich (1972) defined the centrality of an actor based on sum of its adjacent centralities. In our network, researchers who have high eigenvector centrality values will be identified as *leaders* in the co-authorship network since they are connected with too many other influential and highly central researchers, and it is hence expected that they shape the collaborations and play an important role in setting priorities in scientific projects.

Closeness centrality (*cl*) was first proposed by Sabidussi (1966) and is defined based on the shortest path between the nodes in a graph. This measure of centrality considers both direct and indirect connections among the nodes. Researchers with high closeness centrality can be regarded as *local influencers* since they are highly connected to other researchers at least within their own community. These researchers are able to interact and communicate faster (if required) with other researchers in a network hence bringing a brokerage advantage to themselves as well as to their surrounding community. Such researchers are not necessarily important players at the global network level but they are often locally very important actors as they can have an influence on the information spread and on the access to the key resources (both human and financial). Thus, we decided to focus on this role in our analysis as well. Theoretically, the closeness centrality of node i in a graph with n nodes is calculated as:

$$cl_i = \frac{1}{\left(\sum_{j \in n-\{i\}} d(i,j)\right)} \tag{5}$$

where $d(i,j)$ is the length of shortest path between the nodes i and j. Based on the definition, closeness centrality can only be calculated in connected components (completely connected sub-network with no isolated nodes) or graphs since if the graph is not connected the denominator becomes ∞ and as a result the closeness centrality would be zero which is not informative. This assumption has been widely used in different studies (*e.g.* Fleming, King, & Juda, 2007; Uzzi & Spiro, 2005) and is justifiable since the core research activities mainly occur in the largest component (Fatt, Ujum, & Ratnavelu, 2010). We will calculate this centrality measure in the largest connected component of the co-authorship networks. Having defined the dependent variables, the model to be estimated for the network positions of researchers in the reduced form (Model 2) is as follows:

$$\begin{bmatrix} bc_i \\ cc_i \\ ec_i \\ cl_i \end{bmatrix} = \beta_1 \times avgFund3_{i-1} + \beta_2 \times noArt3_{i-1} + \beta_3 \times avgIf3_{i-1} + \beta_4 \times avgCit3_{i-1} + \beta_5 \times dc_i + \beta_6 \times careerAge_i + \beta_7 \times dAcademia_i + \beta_8 \times dProvince_i + \beta_9 \times dFundProgram_i + \alpha_i \quad (6)$$

STATA 12 data analysis and statistical software was used to estimate the models. The independent variables are explained in the next section.

3.3. Independent variables

Funding is recognized as one of the main drivers of collaborative activities. Researchers may use the financial support to get involved in new projects, find new partners and cover the coordination costs among the team members. It enables researchers to better internalize the respective duties among the team members (Ubfal & Maffioli, 2011). Moreover, funding can enable the central (important) researchers in a network to make a good balance between their ongoing collaborative activities and new knowledge creation (Porac *et al.*, 2004). Hence, we considered the average amount of funding that a researcher has received over the past three years ($avgFund3_{i-1}$) in the estimation models. In the literature, three-year (*e.g.* Payne & Siow, 2003) or five-year (*e.g.* Jacob & Lefgren, 2011) time windows have been considered for the funding to take effect. We tested both time windows in our models and found more robust results for the three-year time window.

It is expected that highly productive researchers with high quality works possess more important positions in the network. One reason is that they are expected to work on relatively higher number of projects and as a result get in contact with more researchers. This provides them with more opportunities to get access to new financial and expertise resources which can be used to improve their network position. Therefore, past productivity of researchers in terms of number of publications was also included in the model. Apart from the rate of publications, quality (impact) of works can also play a role in acquiring important positions in a network. High quality works are expected to bring more reputation and distinction to a researcher thus helping him/her to possess more central positions. As a proxy for the quality of the papers, we added $avgIf3_{i-1}$ to the model that was calculated based on the average impact factor of the journals in which the author has published articles in a three year time interval. We also added $avgCit3_{i-1}$ variable to the model that is the average citations of the articles in the past three years as another measure for the impact of the papers. Both measures were included since they reflect slightly different aspects of the publication quality. Journal ranking reflects both the credibility of the journal and the author in a way that it can be sometimes biased toward highly prolific reputable researchers. On the other hand, the citation counts in general reflect the credibility of a paper within the target scientific community.

Degree Centrality (dc) is defined based on the number of ties that a node has (degree) in an undirected graph. Hence, researchers with high degree centrality should be more active since they have higher number of ties (links) to other researchers (Wasserman, 1994). In co-authorship networks it can be regarded as the number of direct collaborators or team members of a researcher thus facilitates access to a variety of skills and complementary knowledge for him/her. Therefore, we expect this measure to play a role in acquiring/maintaining important network positions. Degree centrality for node i is thus defined based on the node's degree and then the values are normalized between 0 and 1 in order to be able to compare centralities:

$$dc_i = \frac{\text{degree of node } i}{\text{highest degree in the network}} \quad (7)$$

Older researchers in general can be more productive (Merton, 1973; Kyvik & Olsen, 2008). Several factors like better access to the funding and expertise sources, more established collaboration network, better access to modern equipment, *etc.* may cause the higher productivity. In addition, senior researchers are on average more recognized than their young counterparts. This recognition might provide a senior researcher with more opportunities to get access to the key knowledge and financial resources and acquire more central positions in his/her scientific community. Hence as a proxy for the career age of the researchers, we included a control variable named $careerAge_i$ representing the time difference between the date of his/her first article in the database and the given year.

In each of the models we used different types of dummy variables. In order to test the impact of the affiliation type we defined $dAcademia_i$ dummy variable indicating whether the researcher is affiliated with academic or non-academic environments. The variable assumes a value of one if a given researcher is affiliated with academia and zero otherwise. This may help to identify which collaboration network roles are more likely to be acquired by industrial and which by academic researchers. We also defined another dummy variable $dProvince_i$ representing Canadian provinces to distinguish the location and geographical impact in possessing various network roles. For example, Quebecers are expected to have on average higher clustering coefficient and to collaborate in more knit groups due to the language factor. To compare the impact of different NSERC funding programs another dummy variable was defined ($dFundProgram_i$). For this purpose, five general classes of NSERC funding programs were considered, *i.e.* discovery grants program, strategic projects, collaborative grants, tools and industrial programs and their impact was assessed on the examined network positions. Discovery grants program was previously named as *research grants program* and is allocated to general long term research activities of the Canadian researchers instead of individual research projects. Since most of the Canadian researchers are being funded by this program, it was considered as the omitted variable. NSERC tool grants support the purchasing of research tools and instruments while industrial programs focus on industrial projects and collaboration. This analysis helps to detect programs that have had more influence on researchers' various network positions. We expect more targeted and high priority programs (*e.g.* strategic projects) to have higher impact on central positions as they can bring some strategic advantages to the researcher. In the next section, the estimation method is presented.

3.4. Method

We used simultaneous multiple regression (Ordinary Least Squares (OLS) method) to estimate both models. In simultaneous approach all the independent variables are treated simultaneously and none of them is considered to be prior to any other ones. As the first stage, a primary list of exploratory variables was prepared including all the candidates of different types. The listed independent variables and the meaningful combinations were added to the model and the results were checked to obtain the final list of independent variables (introduced in the previous section) for which the method produced the most significant and robust results and the OLS assumptions were satisfied. This included the checks for the normality of the residuals, controlling for homoscedasticity of errors to ensure the homogeneity of variance of the residuals and existence of multicollinearity among the independent variables. To test the severity of multicollinearity, we did the variance inflation factor (VIF) diagnostic test. According to the test results, non-existence of simultaneity which is an important source of endogeneity was guaranteed. To check the linearity assumption, we plotted each of the independent variables against the standardized residuals to make sure that there was no clear sign of nonlinearity, thus assuring that the plots were just a random distribution of points. The correlations among the variables were also considered in selecting the final variables.

In addition, we log-transformed the variables for which their distribution was not normal. We also checked for bivariate scatterplot of the exploratory variables of interest to decide if it is needed to add a non-linear component. This yielded to the inclusion of the square term of the *careerAge* variable in some of the models that enabled us to see the curvature of the relationship. A high VIF was just observed between *careegAge* and $careerAge^2$ variables which was expected but not of an issue since they are not really two different variables. We also included the interaction variable of degree centrality and career age in Model 2 as we expected a lower number of direct collaborators after a certain age. Finally, we tested the models for the robustness and consistency of the results. In addition, we did bootstrap re-sampling to verify the stability of the models by bootstrapping the standard errors of the parameter estimates by performing 50 replications. The intensive preprocessing stages that were explained earlier helped us to collect an accurate and complete data. This clean large dataset, the inclusion of several independent variables of various types along with the intensive tests contributed to obtain satisfactory models.

4. Results

4.1. Descriptive analysis

Before turning to the regression models, we first analyze the overall trends of the dependent variables as well as funding, as the main determinant influencing factor of scientific activities (Martin, 2003). Fig. 3 presents the average amount of NSERC funding per researcher during the examined time interval. As it can be seen average funding received per researcher has been following an increasing trend while after 2003 (solid vertical line in Fig. 3) the slope has become steeper indicating a considerable increase in the average amount of NSERC funding. In addition, during the first five years of the examined time interval (dashed vertical line in Fig. 3) we see a steadier trend of average funding in comparison with the other periods. We will use the vertical lines of average funding in the rest of the figures of this section to assess the impact of funding easier. In addition, in the rest of the paper *funding period I*, *II*, and *III* will refer to the periods of 1996–2000, 2000–2003, and 2003–2010 respectively.

Researchers publish their results in books or journal articles or present them at scientific conferences in order to preserve priority for their discoveries and raise their scientific reputation. Although most of the articles were single authored till 1920s (Greene, 2007), today in most of the academic disciplines (except humanities) researchers prefer multi-authorship model due to the nature of the big science that requires collaboration and expertise of many individuals (De Solla Price, 1986). Number of authors per paper has been considered as a proxy for scientific collaboration in several studies (*e.g.* Newman,

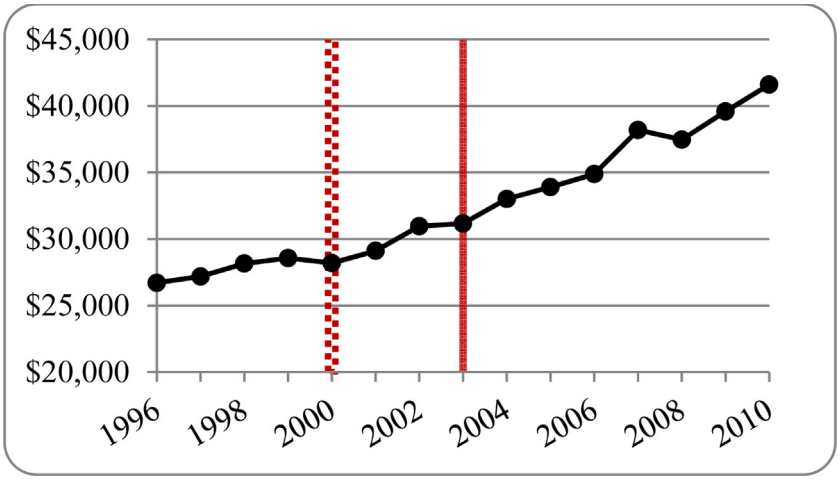

Fig. 3. Average funding per researcher.

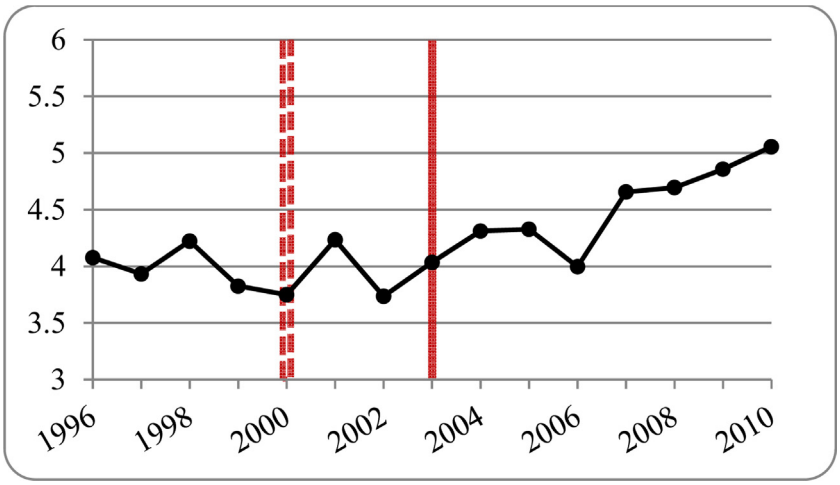

Fig. 4. Average number of authors per article.

2004; Rosenzweig *et al.*, 2008). Fig. 4 presents the average number of authors per paper for the NSERC funded researchers. The vertical lines show different periods of average funding that were discussed earlier. According to Fig. 4, as the amount of average funding increases the average number of authors per articles also augments. In other words, it seems funding can be regarded as one of the influencing factors on the researchers' team size as higher amount of funding might enable researchers to form larger scientific teams aiming to increase their productivity. This is quite reasonable since apart from the higher complexity of science the competition among scientists to get access to better resources has also increased, hence the average number of authors per paper is augmenting (Powers, 1988).

Trends of the network structure variables are represented in Fig. 5. As it can be seen, clustering coefficient of the co-authorship networks is steady during the whole time interval. Except some minor jumps, the overall average trend of degree centrality is also almost steady. However, a considerable decline in degree centrality is observed during the years of the funding period I. Although the trend of betweenness centrality is steady during the funding period I, it drastically increases within the funding period II maintaining its level in funding period III despite some fluctuations. Hence, according to Figs. 3–5 it seems that at the aggregate level there is a positive relation between funding and collaboration measured by average number of authors per article. However, nothing can be said about the network structure variables. Hence, in order to assess the effects more accurately we turn to the regression analysis to investigate the impact of the influencing factors on collaboration at the individual level.

4.2. Statistical analysis

As discussed in Section 3.2, we have two types of dependent variables, one is the more common type of collaboration indicator measured by the average number of authors per paper, and the other one is based on the network structure

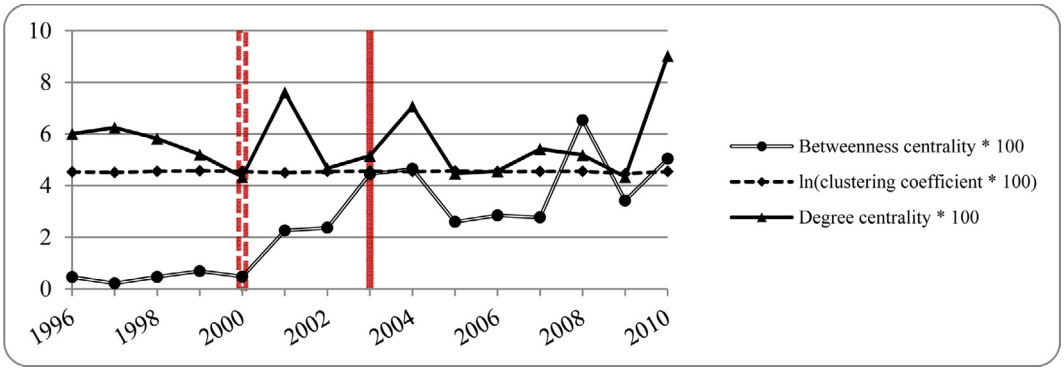

Fig. 5. Average betweenness centrality, clustering coefficient, and degree centrality per year.
We did not include closeness centrality in the figure as it is calculated in the largest component while the other network measures are calculated for the whole network.

variables. In this section, the regression results are presented and discussed for both types of the dependent variables. The correlation matrices for all the regression models are presented in Appendix A.

4.2.1. Average number of authors per paper (Model 1)

We analyzed the impact of the influencing factors on the scientific team size of the researchers measured by average number of authors per article at the individual level. The scientific team size was calculated in two ways, one by considering all the co-authors of a researcher (Model 1-a) and the other one by taking the distinct co-authors into the account (Model 1-b). Distinct team size of a researcher was calculated by counting the distinct number of his/her co-authors divided by the number of his/her publications. In the overall team size model, the numerator is the overall number of co-authors for a researcher. For all the models, we considered all the combinations of the lags for the variables in the model and used the ones that yielded the most robust results. This is similar to the approach of Schilling and Phelps (2007) and Beaudry and Allaoui (2012). The regression results are presented in Table 1.

In Model 1-a, as it can be seen the average amount of researcher's funding in the past three years has a significant and relatively high positive impact on overall team size of the researcher. This is in accordance with several studies (*e.g.* Adams *et al.*, 2005; Gulbrandsen & Smeby, 2005) who found that larger amount of funding will positively affect the scientific collaboration. As expected, the past productivity of the funded researcher measured by the number of articles over a three-year time window (*noArt3*) has also a positive impact on the team size. This may partially highlight the importance of collaboration in scientific activities in a way that highly productive researchers may benefit from larger scientific teams. According to the results not only the rate of publications affects the team size, the impact of the works also positively influences the collaboration (*avgCit3* and *avgIf3*). In other words, higher quality papers of the NSERC funded researchers in the past three-year have a positive relation with their scientific team size in the following year. Hence, the results suggest that productive researchers with high quality works are more collaborative.

We controlled for the age of the researchers in the regression model and it was observed that the career age of the funded researchers negatively influences their collaboration. Despite the advantages of collaboration (*e.g.* better access to resources, internal referring, *etc.*), there are some costs (*e.g.* finding right partners and research coordination) related to the scientific collaboration (He, Geng, & Campbell-Hunt, 2009). As an example, Cummings and Kiesler (2007) focused on the effects of the coordination costs on collaboration among U.S. universities and found that coordination failures have a negative impact on scientific collaboration. Hence, it seems that as the career age of the researchers grows negative impact of costs of collaboration increases in a way that at a certain level senior researchers may tend not to increase their team size.

Table 1
Regression results, team size models (Model 1).

	Model 1-a Overall team size	Model 1-b Distinct team size
$ln\_avgFund3_{i-1}$	1.092 (.212)***	.835 (.133)***
$noArt3_{i-1}$	.207 (.041)***	.047 (.025)*
$ln\_avgCit3_{i-1}$	1.19 (.194)***	.604 (.122)***
ln_avgIf3_{i-1}	4.354 (.317)***	2.461 (.197)***
$careerAge_i$	−.712 (.218)***	−.435 (.137)***
$careerAge^2{}_i$	.035 (.013)***	.024 (.008)***
Affiliations dummy variable		
Academia	−8.097 (1.126)***	−4.633 (.706)***
Number of observations	60907	60907

Notes: Standard errors in parentheses.
Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 2
Regression results, network structure models (Model 2).

	Model 2-a Betweenness centrality ($bc^4{}_i$)	Model 2-b Clustering coefficient (cc_i)	Model 2-c Eigenvector centrality ($ec^4{}_i$)	Model 2-d Closeness Centrality ($cl^2{}_i$)
$ln\_avgFund3_{i-1}$	.436 (.066)***	−.007 (.002)***	−.397 (.079)***	.169 (.027)***
$noArt3_{i-1}$	.469 (.011)***	−.013 (.0004)***	.048 (.013)***	.022 (.003)***
$ln\_avgCit3_{i-1}$	.939 (.061)***	.029 (.002)***	.186 (.074)**	−.073 (.026)***
ln_avgIf3_{i-1}	−.288 (.103)***	.013 (.003)***	.061 (.124)	.475 (.044)***
$dc^4{}_i$	.01 (.002)***	.001 (.0001)***	.307 (.005)***	.026 (.0005)***
$careerAge_i$	−.033 (.017)*	−.012 (.002)***	.019 (.02)	−.419 (.026)***
$careerAge^2{}_i$		.001 (.0001)***		.024 (.024)***
Interaction variable				
$dc^4{}_i \times careerAge_i$		−.00004 (.00002)**	−.002 (.0006)***	
Affiliations dummy variable				
Academia	−.043 (.399)	−.129 (.014)***	1.184 (.483)**	.06 (.144)
Provinces dummy variables				
Quebec		.035 (.005)***		.145 (.059)**
British Columbia		.006 (.006)		.197 (.069)***
Alberta		−.017 (.006)***		.239 (.072)***
Saskatchewan		.015 (.011)		.261 (.121)**
New Brunswick		−.022 (.014)		.118 (.17)
Manitoba		.006 (.011)		.275 (.138)**
Newfoundland		−.001 (.015)		.169 (.19)
Prince Edward		.047 (.035)		−.169 (.52)
Nova Scotia		.018 (.01)*		.155 (.116)
Funding programs dummy variables				
Strategic		.031 (.008)***		.076 (.085)
Tools		.041 (.011)***		.759 (.126)***
Collaborative		.038 (.008)***		.068 (.086)
Industrial		−.01 (.013)		−.293 (.134)**
Number of observations	38,974	38,974	38,974	15,046

Notes: Standard errors in parentheses.
Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We also added a quadratic term of the career age ($careerAge^2$) to see the curvature of the relationship and realized that the curve of the career age is convex, *i.e.* a sharp increase in team size becomes steady after a certain age. To assess the impact of the type of the affiliation of the researcher on collaboration, the institution type dummy variable (*dAcademia*) that takes value 1 if the funded researcher belongs to the academic environment and 0 if his affiliation is non-academic was also added to Model 1-a. As it can be seen, academic researchers are significantly different from the non-academic ones and they work in smaller scientific teams in comparison with their non-academic counterparts.

We did the same analysis on the distinct average team size of the funded researchers (Model 1-b). According to the results the signs of the influencing factors are the same as the ones for Model 1-a but the coefficients are smaller indicating a lower intensity of the considered factors. Hence, in general the discussion presented for Model 1-a is also valid for Model 1-b. As it can be seen in Table 1, interestingly the impact factor has the largest coefficient in the both models followed by citation counts in Model 1-a. This highlights the importance of collaborative activities in producing publications of higher quality. Although we expected funding to be the most influencing factor, it came after the publications' quality proxies. According to the results, career age and rate of past publications were found to be the least influencing factors in the both models.

4.2.2. Network structure variables

In this section, the impact of influencing factors on the network structure variables is assessed. Betweenness centrality (Model 2-a), clustering coefficient (Model 2-b), eigenvector centrality (Model 2-c) and closeness centrality (Model 2-d) are considered as the dependent variables. The multiple regression analysis is done at the individual level of the researchers. Regression results are presented in Table 2.

According to the results for Model 2-a, the rate and quality (measured by the average number of citations) of the researchers' papers in the past three years have the highest positive impact on their betweenness centrality in the following year. Hence, it can be said that a researcher with a high number of articles that are on average of high quality possesses a more central position in the co-authorship network, acting as an influential intermediary in knowledge diffusion and the formation of scientific collaboration. In addition, as it was expected the average amount of funding received in the past three years has also a positive impact on the centrality of the funded researcher in a way that more funded researchers would be more probable candidates for the central positions of the network. This finding is partially supported by the small positive impact of the direct number of partners of the researchers measured by their degree centrality (*dc*), since higher amounts of funding may enable researchers to expand their scientific activities that might result in more central positions. Surprisingly,

a negative relation is found between the average impact factor of the journals in which the researchers have published their articles (*avgIf3*) and the betweenness centrality. It seems that the average number of citations is a better proxy for evaluating the quality of the works in the co-authorship network of the NSERC funded researchers. According to the results publishing in higher quality journals will not necessarily lead the researcher to a more influential gatekeeper position. Career age of the researchers has also a negative impact on their betweenness centrality in our examined co-authorship network indicating that as time passes from the date of the first publication of a researcher, betweenness centrality declines. This might be due to the creation of new knowledge brokers in the network. In other words, as the time passes other younger prolific researchers also become more central through getting access to the pools of knowledge in diverse clusters and communities. This might lead to the creation of new knowledge brokers which might decay the role of old brokers. In addition, the affiliation of researchers does not differently affect their central positions and there is no correlation between the type of the affiliation of the researchers and their betweenness centrality. We did the same analysis for the impact of the location of the researchers categorized by different Canadian provinces and found no impact of location on betweenness centrality of researchers.

The next dependent network structure variable was the clustering coefficient (*cc*) of researchers at the individual level (Model 2-b). Clustering coefficient of a researcher in the network indicates the likelihood that two researchers (authors in co-authorship networks) who are connected to a specific third scientist are also connected to one another, forming a clique together, *i.e.* a subset of the nodes of a graph (network) in a way that every two nodes are connected by an edge. In other words, clustering coefficient of a node in a network indicates the ratio of the number of triangles that passes through that node over the maximum number of possible triangles around that node. Hence, clustering coefficient is zero for the nodes with less than two neighbors. According to the results, funding has a very small negative impact on the clustering coefficient of researchers. It can be said that researchers may use the allocated funding to find more new partners rather than making connections among their previous partners to form a triangle. Hence, it seems that more funding will result more in linear expansion of the team size of the researchers. In addition, it is observed that past productivity measured by the number of articles in the previous years has also a negative impact on the clustering coefficient. One reason could be that researchers that are highly productive may have less time to organize and expand the internal connections among their directly connected partners (nodes). Both of the proxies for the quality of the papers (*avgCit3* and *avgIf3*) have a positive impact on the clustering coefficient of the researcher where citation counts measure has the largest coefficient. The nature of the science has become more inter-disciplinary that needs more involvement of researchers from different backgrounds. Hence, it seems that the production of higher quality papers requires more internal communities around a researcher (node) in the form of triangles that could be formed by the involvement of researchers from different disciplines. This might lead to higher clustering coefficient of the researcher. Since the impact of the degree of the node (*dc*) on the clustering coefficient is positive it can be said that in the local network of the researchers with more directly connected partners forming more triangles is more probable that will result in higher clustering coefficient. We added the quadratic form of the career age ($careerAge^2$) in order to see the curvature of the relationship between the career age of researchers and their clustering coefficient. According to the results, although at first the impact of the career age is negative, approximately after 17 years from the first publication of a researcher the overall impact of the career age becomes positive and the clustering coefficient starts to increase. Hence, the curve of the career age is convex with the minimum around the age of 17. Therefore, it can be said that in general mid-career scientists have higher clustering coefficient, which is quite expected since on average they benefit from better established co-authorship and collaboration networks. Since a negative effect is observed for the interaction variable of the career age and degree of a node ($dc \times careerAge$), it can be said that there is a balance between the number of direct partners of a researcher and his/her age. In other words, although it was found that mid-career scientists are on average more cliquish, if they have too many direct partners it may affect their cliquishness negatively. In general, the estimated coefficients are small in comparison with other models indicating the lower influence of the factors on clustering coefficient. The analysis of the institution type dummy variable reveals that academic researchers are significantly different from the non-academic ones and have on average around 13% (−.129) less cliquishness in comparison with the non-academic researchers. We considered Ontario as the omitted dummy variable in the analysis of the Canadian provinces. According to the results, researchers who are located in Quebec, Nova Scotia, and Alberta are significantly different from the ones who reside in Ontario. However, the coefficient is positive only for Quebec and Nova Scotia, which may indicate higher clustering coefficient of the researchers located in the mentioned provinces. As explained before, we defined dummy variables for the most frequent NSERC funding programs, namely discovery grants, strategic projects, industrial funding, collaborative grants, and tools and equipment grants. The dummy variable of the discovery grants was omitted. According to the results, it can be seen that the effects of strategic, tools, and collaborative funding programs are significantly and positively different from the discovery grants program (the omitted variable). These findings were expected specifically for the strategic project grants which have the highest coefficient among the mentioned programs. Based on the definition of the strategic project funding programs, the aim is to improve the scientific development in selected high-priority areas that influence Canada's economic and societal position. Hence, these well-defined targeted grants should be allocated to specific reputable researchers who might possess more central positions in the network according to the regression analysis.

According to the results of Model 2-c, funding (*avgFund3*) has surprisingly a negative impact on eigenvector centrality. Hence, it seems that higher funding may reduce the leadership role, possibly by involving the highly funded researchers in other scientific activities like defining new projects, finding new partners, *etc.* It can be seen that the average journal

impact factor and the career age of the researchers do not have a significant impact on researchers' eigenvector centrality. However, past productivity of the researchers in terms of both quantity (*noArt3*) and quality (*avgCit3*) of the papers has a positive impact. The reason could be that being more productive may increase the chance of meeting/cooperating with other reputable productive researchers who possess central positions in the network. The degree centrality of a node, as a measure of the direct number of partners of a scientist, has also a relatively large positive effect on eigenvector centrality. It was quite expected since researchers with high eigenvector centrality should have high number of connections from which most of the connections would be high-profile central scientists. However, researchers who have high eigenvector centrality (named as leaders) do not necessarily occupy positions with high betweenness centrality (acting as gatekeepers) or even high closeness centrality (acting as local influencers). But, they are highly connected with mainly high profile individuals within highly interconnected clusters. Interestingly, the interaction of degree and career age of the researchers shows a negative effect on eigenvector centrality. This might indicate that as the career age of the researchers grows, higher number of direct connections may affect their leadership role negatively. Of course, there should exist a balance between age, degree, and eigenvector centrality. The analysis of the institution type dummy variable (*dAcademia*) reveals that academic researchers are significantly different from their industrial counterparts. The positive coefficient of the dummy variable indicates that academic researchers are more likely to have higher eigenvector centrality (to act as leaders) in the co-authorship networks rather than the non-academic scientists. In addition, the large coefficient of the institution type dummy variable highlights the more important role of affiliation in having higher eigenvector centrality in comparison with other selected network variables.

Finally, we discuss the impact of influencing factors on closeness centrality of researchers (Model 2-d). According to the results, average funding (*avgFund3*) positively affects closeness centrality of the researchers. Hence, it can be said that more funding may enable researchers with high closeness centrality (who are important influencers within their local network) to increase their penetration and prestige. Although a small positive effect was observed for the rate of publication (*noArt3*) on the closeness centrality, the relation between the quality of the papers and closeness centrality is not very clear, since the citation based proxy (*avgCit3*) shows a negative impact while the journal impact factor based measure (*avgIf3*) presents a larger positive effect. Hence, it seems that local influencers do not necessarily publish high quality works. As it was expected, the direct number of partners of the researchers, measured by degree centrality (*dc*), has a significant large positive impact on closeness centrality since local influencers may benefit from larger team sizes and higher number of connections to empower their penetration within their local community. The quadratic term of the career age ($careerAge^2$) was also added to the model to investigate the curvature of the relationship. Based on the results, the impact of the career age on closeness centrality of the researchers is negative at first. However, approximately after 18 years the overall impact of the career age becomes positive. Therefore, the curve of the career age in the closeness centrality model is convex with the maximum around the age of 18. Hence, it seems that mid-career scientists are more likely to have higher influence within their local community. As it can be seen academic and non-academic researchers (measured by *dAcademia*) do not have significantly different impact on the closeness centrality. Hence, it is equally likely that local influencers come from industry or academic environments. In addition, researchers who are located in Quebec, British Columbia, Alberta, Saskatchewan, and Manitoba are significantly different from the ones who reside in Ontario. The coefficient is positive for all the mentioned provinces indicating higher closeness centrality of the researchers located in the mentioned provinces in comparison with their counterparts in Ontario. The coefficient was the highest for the researchers who reside in Manitoba. The analysis of different NSERC funding programs shows that the effect is only different for tools and industrial funding programs, with positive and negative coefficients respectively.

5. Conclusion

In this paper we investigated the impact of funding and other influencing factors like past productivity, number of direct scientific partners, and career age of the researchers on their positions and roles within the co-authorship networks. We employed social network analysis and statistical approaches to assess the impact of the mentioned factors on the network structure variables. We did the analysis both for the common indicators of scientific collaboration that are based on the number of authors per paper and for the network structure variables. To our knowledge this is the first study that considers the network structure measures as dependent variables and analyzes various factors which affect them at the individual level.

Analyzing the impact of the influencing factors on the traditional collaboration and scientific team size indicators revealed that funding plays a significant positive role in motivating researchers to collaborate more. This finding is in line with several studies, *e.g.* Adams *et al.* (2005) and Gulbrandsen and Smeby (2005). In addition, it was observed that highly prolific researchers who are producing high quality papers have on average larger scientific teams. This partially confirms the importance of collaboration in scientific activities. Analyzing the career age of the researchers showed that it negatively influences their collaboration, which might be partially due to difficulties in managing the costs of collaboration (*e.g.* finding right partners and research coordination). Another reason can be higher motivation of young researchers to find new partners and get involved in new projects in order to improve their academic position and secure more funding.

In the second part of the analysis the impact on the network structure variables was investigated. Researchers with high betweenness centrality (gatekeepers) are often critical to scientific collaboration and knowledge diffusion as they

can control the flow of information and collaboration. Our results suggest that the past productivity of the researchers in terms of both quantity and quality of the publications along with the average amount of funding available are crucial factors in achieving higher betweenness centrality. Analyzing the impact of degree centrality as a measure of the number of direct partners of a researcher on the betweenness centrality revealed that in the examined co-authorship network higher number of direct connections empowers the role of gatekeepers. Surprisingly, a negative impact of the career age of the researchers on their betweenness centrality was observed. This might indicate the considerable role of young gatekeepers in connecting different scientific communities (clusters) and knowledge diffusion in the examined collaboration network.

Researchers with high clustering coefficient (cliquishness) are the ones who prefer to collaborate in *knit groups*. According to the results, funding has a negative impact on knit group collaboration. This might indicate the linear use of funding resources by researchers in the examined network leads to the expansion of their direct partners rather than empowering their internal teams through formation of triangles among researchers. Interestingly, a negative impact of the rate of publication on the cliquishness was observed, while the impact of the quality of the papers was positive. This might partially highlight the role of interdisciplinary research in a way that higher quality publications cause the formation of more triangles among the researchers. On the other hand, knit group collaborators may form internal scientific communities (teams) in order to increase the quality of their work though *e.g.* having reviewed their works by several experts (internal referring). Analyzing the effect of career age revealed that approximately after 17 years from the date of researchers' first publication, the overall impact of the career age on the clustering coefficient becomes positive. Hence, in general, mid-career scientists tend to work in denser local scientific communities (higher clustering coefficient) which is quite expected since on average they benefit from better established co-authorship and collaboration networks. However, although they engage more in a dense local collaboration, the number of their direct connections will be affected by their career age in a way that if they continue to increase the number of their direct partners their influence on their local community decreases after several years.

Analyzing the eigenvector centrality has been mostly neglected in the studies that assessed co-authorship networks. Researchers with high eigenvector centrality can be identified as the leaders among their connections since they have often many connections to reputable highly central researchers. Therefore, they can play an important role in forming scientific collaboration teams or in defining new projects and setting priorities in the projects. Surprisingly, a negative impact of funding on the eigenvector centrality was observed. This might indicate that higher funding may reduce the leadership role, possibly by involving the highly funded researcher in other scientific activities like defining new projects, finding new partners, *etc*. Moreover, past productivity of the researchers in terms of both quantity and quality of the papers has a positive impact on their leadership role that is quite expected. One reason could be that being more productive may increases the chance of meeting/cooperating with other reputable productive researchers who possess central positions in the network. This finding was also confirmed by the positive impact of the degree centrality on the eigenvector centrality since higher number of direct connections increases the probability of meeting/cooperating with high-profile central scientists that will result in higher eigenvector centrality. However, since the interaction of degree and career age of the researchers presents a negative effect on eigenvector centrality it might be suggested that as the career age of the researchers grows, higher number of direct connections may affect their leadership role negatively.

Finally, we assessed the impact of the influencing factors on the closeness centrality of the researchers in the largest connected components of the co-authorship networks and at the individual level. Researchers with high closeness centrality are identified as important local influencers within their local collaboration network or community. Although they might not be important actors in the entire network, they are highly respected locally as they are on the local short paths of knowledge diffusion. Our results showed a positive impact of funding on the closeness centrality suggesting that local influencers may use more funding to increase their penetration and prestige within their local community. Analyzing the impact of past productivity revealed that local influencers are not necessarily highly prolific scientists, especially in terms of the quality of their publications. However, number of direct connections plays an important role in a way that local influencers can use it to empower their penetration within their local community. Analyzing the impact of the career age showed that the overall career age impact becomes positive after 18 years hence it seems that mid-career scientists are more likely to have higher influence within their local community.

According to the results, some implications in terms of public policy can be made. First, collaboration networks are evolving all the time. In addition, nature of the science is also becoming more inter-disciplinary, thus in order to facilitate the formation of efficient and effective dynamic collaboration networks the most important (influential) researchers are expected to be the most central ones. It is specifically important for the ones who perform the leadership role. Hence, it can be suggested to support the central scientists to foster the scientific collaboration as such researchers are vital for facilitating the flow of knowledge and information. In addition, they can play an important role in connecting other researchers/communities thus making a tighter collaborative environment. This will also help researchers who are connected to highly central researchers to have on average a better access to the expertise and financial resources. And, it is also expected that prolific highly influential experienced researchers better manage the collaboration costs, hence supporting them might result in a more efficient network. Second recommendation is related to the role of young researchers in knowledge transfer and flow control who were found to play an important role in connecting different scientific communities and in knowledge diffusion through their collaboration network. Moreover, it seems that the allocation of funding is now

biased toward the senior researchers that helps them to occupy more central positions. These results thus support an establishment of a policy which would encourage young researchers to take gatekeeper roles, for example by providing them with increased financial resources. In addition, as it was observed the rate of collaboration decreases with the career age. Therefore, supporting young prolific researchers can not only foster the scientific collaboration, but it can also deliver new reliable leaders for future. Last but not least, as it was observed less productive industrial researchers prefer to work in knit groups. Hence, it would be suggested to encourage academia-industry collaboration more by involving highly productive academic researchers in industrial projects.

6. Limitations and future work

We were exposed to some limitations in this paper. First, we selected SCOPUS for gathering information about the NSERC funded researchers' articles. Since SCOPUS and other similar databases are English biased, hence, non-English articles are underrepresented (Okubo, 1997). Second, since SCOPUS data was less complete before 1996, we chose the time interval of 1996 to 2010 for our analysis. Another inevitable limitation about the data was the spelling errors and missing values. Although SCOPUS is confirmed in the literature to have a good coverage of articles, as a future work it would be recommended to focus on other similar databases to compare and confirm the results.

We measured closeness centrality in the largest component of the co-authorship networks since based on the classic definition of the closeness centrality it can be defined in connected graphs or sub-graphs. Some other approaches for calculating the closeness centrality in disconnected graphs have been proposed in the literature (*e.g.* Latora & Marchiori, 2001; Dangalchev, 2006). However, there are still doubts about such new approaches to be counted as extensions of the closeness centrality (Yang & Zhuhadar, 2011). Future works can address this issue by considering the new approaches and comparing the results with the ones of the classic method of calculation of closeness centrality.

Appendix A. Correlation matrices

See Tables A1–A6

Table A1
Correlation matrix, overall team size model.

Variable	$teamSize_i$	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$careerAge_i$
$teamSize_i$	1.0000					
$ln\_avgFund3_{i-1}$	.0341	1.0000				
$noArt3_{i-1}$	.0314	.4229	1.0000			
$ln\_avgCit3_{i-1}$	.0552	.1207	.0742	1.0000		
ln_avgIf3_{i-1}	.0764	.1081	.0481	.4014	1.0000	
$careerAge_i$	.0039	.3201	.2940	.2047	.0228	1.0000

Table A2
Correlation matrix, distinct team size model.

Variable	$teamSizeDis_i$	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$careerAge_i$
$teamSizeDis_i$	1.0000					
$ln\_avgFund3_{i-1}$	.0338	1.0000				
$noArt3_{i-1}$	.0199	.4229	1.0000			
$ln\_avgCit3_{i-1}$	.0481	.1207	.0742	1.0000		
ln_avgIf3_{i-1}	.0683	.1081	.0481	.4014	1.0000	
$careerAge_i$	.0056	.3201	.2940	.2047	.0228	1.0000

Table A3
Correlation matrix, betweenness (bc) model.

Variable	bc_i	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	dc_i	$careerAge_i$
bc_i	1.0000						
$ln\_avgFund3_{i-1}$	.1467	1.0000					
$noArt3_{i-1}$	.2559	.4403	1.0000				
$ln\_avgCit3_{i-1}$	.1031	.1206	.0899	1.0000			
ln_avgIf3_{i-1}	.0386	.1197	.0547	.4037	1.0000		
dc_i	.0394	.0459	.0522	.0586	.1114	1.0000	
$careerAge_i$	.0930	.3406	.3062	.2361	.0251	−.0193	1.0000

Table A4
Correlation matrix, clustering coefficient (*cc*) model.

Variable	cc_i	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$dc^*10^4{}_i$	$careerAge_i$
cc_i	1.0000						
$ln\_avgFund3_{i-1}$	−.0982	1.0000					
$noArt3_{i-1}$	−.2018	.4403	1.0000				
$ln\_avgCit3_{i-1}$	.0746	.1206	.0899	1.0000			
ln_avgIf3_{i-1}	.0496	.1197	.0547	.4037	1.0000		
$dc \times 10^4{}_i$	.0571	.0459	.0522	.0586	.1114	1.0000	
$careerAge_i$	−.0686	.3406	.3062	.2361	.0251	−.0193	1.0000

Table A5
Correlation matrix, eigenvector centrality (*ec*) model.

Variable	ec_i	$ln\_avgFund3_{i-1}$	$noArt3(_{i-1}$	$ln\_avgCit3(_{i-1}$	$ln\_avgIf3(_{i-1}$	dc_i	$careerAge_i$
ec_i	1.0000						
$ln\_avgFund3(_{i-1}$	.0088	1.0000					
$noArt3(_{i-1}$	.0353	.4403	1.0000				
$ln\_avgCit3(_{i-1}$	.0412	.1206	.0899	1.0000			
$ln\_avgIf3(_{i-1}$	.0604	.1197	.0547	.4037	1.0000		
dc_i	.4916	.0459	.0522	.0586	.1114	1.0000	
$careerAge_i$	−.0059	.3406	.3062	.2361	.0251	−.0193	1.0000

Table A6
Correlation matrix, closeness centrality (*cl*) model.

Variable	cl_i	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$dc \times 10^2{}_i$	$careerAge_i$
cl_i	1.0000						
$ln\_avgFund3_{i-1}$	.0821	1.0000					
$noArt3_{i-1}$	.0806	.4656	1.0000				
$ln\_avgCit3_{i-1}$	.0490	.0936	.0873	1.0000			
ln_avgIf3_{i-1}	.1574	.1013	.0501	.4421	1.0000		
$dc \times 10^2{}_i$	.3845	.0329	.0318	.0496	.1619	1.0000	
$careerAge_i$	−.144	.3808	.3175	.0768	−.0285	−.0493	1.0000

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*(4), 594–607.

Abbasi, A., Altmann, J., & Hwang, J. (2010). Evaluating scholars based on their academic collaboration activities: Two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, *83*(1), 1–13.

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403–412.

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from US universities, 1981–1999. *Research Policy*, *34*(3), 259–285.

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

Balconi, M., & Laboranti, A. (2006). University–industry interactions in applied research: The case of microelectronics. *Research Policy*, *35*(10), 1616–1630.

Beaudry, C., & Allaoui, S. (2012). Impact of public and private research funding on scientific production: The case of nanotechnology. *Research Policy*, *41*(9), 1589–1606.

Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration—Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799–1830. *Scientometrics*, *1*(2), 133–149.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*(1), 113–120.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy*, *33*(4), 599–616.

Brad Wray, K. (2006). Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science, A*, *37*(3), 505–514.

Cronin, B. (2005). *The hand of science: Academic writing and its rewards*. Lanham, MD: Scarecrow Press.

Cummings, J. N., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, *36*(10), 1620–1634.

Dangalchev, C. (2006). Residual closeness in networks. *Physica A: Statistical Mechanics and its Applications*, *365*(2), 556–564.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek (Structural analysis in the social sciences)*. , illustrated edition.

De Solla Price, D. J. (1963). *Big science, little science*. New York, NY: Columbia University.

De Solla Price, D. J. (1986). *Little science, big science... and beyond*. New York, NY: Columbia University Press.

Defazio, D., Lockett, A., & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy*, *38*(2), 293–305.

Ebadi, A., & Schiffauerova, A. (2013). Impact of funding on scientific output and collaboration: A survey of literature. *Journal of Information & Knowledge Management*, *12*(04).

Ebadi, A., & Schiffauerova, A. (2015a). On the relation between the small world structure and scientific activities. *PLoS ONE*, *10*(3), e0121129. http://dx.doi.org/10.1371/journal.pone.0121129

Ebadi, A., & Schiffauerova, A. (2015b). How to receive more funding for your research? Get connected to the right people!. *PLoS ONE*, *10*(7), e0133061. http://dx.doi.org/10.1371/journal.pone.0133061

Eslami, H., Ebadi, A., & Schiffauerova, A. (2013). Effect of collaboration network structure on knowledge creation and technological performance: The case of biotechnology in Canada. *Scientometrics*, *97*(1), 99–119.

Fatt, C. K., Ujum, E. A., & Ratnavelu, K. (2010). The structure of collaboration in the Journal of Finance. *Scientometrics*, *85*(3), 849–860.

Fleming, L., King, C., III, & Juda, A. I. (2007). Small worlds and regional innovation. *Organization Science*, *18*(6), 938–954.
Godin, B. (2003). *The impact of research grants on the productivity and quality of scientific research. No. 2003. INRS working paper*.
Greene, M. (2007). The demise of the lone author. *Nature*, *450*(7173), 1165–1165.
Grossman, J. W. (2002). The evolution of the mathematical research collaboration graph. In *Congressus Numerantium*.
Gulbrandsen, M., & Smeby, J. C. (2005). Industry funding and university professors' research performance. *Research Policy*, *34*(6), 932–950.
Håkansson, H., & Ford, D. (2002). How should companies interact in business networks? *Journal of Business Research*, *55*(2), 133–139.
Hanneman, R. A., & Riddle, M. (2011). Concepts and measures for basic network analysis. In *The Sage handbook of social network analysis*.
He, Z., Geng, X., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy*, *38*(2), 306–317.
Huang, J., Zhuang, Z., Li, J., & Giles, C. L. (2008). Collaboration over time: Characterizing and modeling network evolution. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 107–116).
Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9), 1168–1177.
Jiang, Y. (2008). Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, *74*(3), 471–482.
Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.
Kyvik, S., & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics*, *76*(3), 439–455.
Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters*, *87*(19), 198701.
Martin, B. R. (2003). The changing social contract for science and the evolution of the university. In *Science and innovation: Rethinking the rationales for funding and governance*. Cheltenham: Edward Elgar.
Mattsson, L., & Johanson, J. (1992). *Network positions and strategic action: An analytical framework*. Univ.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.
Mote, J. E. (2005). R&D ecology: Using 2-mode network analysis to explore complexity in R&D environments. *Journal of Engineering and Technology Management*, *22*(1), 93–111.
Newman, M. E. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. *Complex Networks*, *650*, 337–370.
Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: Methods and examples, OECD, science, technology and industry. Working papers, no. 1997/01* Paris: OECD Publishing.
Payne, A. A., & Siow, A. (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy*, *3*(1), 1–22.
Porac, J. F., Wade, J. B., Fischer, H. M., Brown, J., Kanfer, A., & Bowker, G. (2004). Human capital heterogeneity, collaborative relationships, and publication patterns in a multidisciplinary scientific alliance: A comparative case study of two scientific teams. *Research Policy*, *33*(4), 661–678.
Powers, R. D. (1988). Multiple authorship, basic research, and other trends in the emergency medicine literature (1975 to 1986). *The American Journal of Emergency Medicine*, *6*(6), 647–650.
Rosenzweig, J. S., Van Deusen, S. K., Okpara, O., Datillo, P. A., Briggs, W. M., & Birkhahn, R. H. (2008). Authorship, collaboration, and predictors of extramural funding in the emergency medicine literature. *The American Journal of Emergency Medicine*, *26*(1), 5–9.
Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.
Schilling, M. A., & Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, *53*(7), 1113–1126.
Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.
Tijssen, R. J. (2004). Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. *Research Policy*, *33*(5), 709–733.
Tijssen, R. J., van Leeuwen, T. N., & Korevaar, J. C. (1996). Scientific publication activity of industry in the Netherlands. *Research Evaluation*, *6*(2), 105–119.
Ubfal, D., & Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, *40*(9), 1269–1279.
Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, *111*(2), 447–504.
Wasserman, S. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.
Yang, R., & Zhuhadar, L. (2011). Extensions of closeness centrality? In *Proceedings of the 49th annual southeast regional conference* (pp. 304–305).