Impact of Funding on Scientific Output and Collaboration

Ashkan Ebadi

A Thesis

In the Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Information Systems Engineering) at

Concordia University

Montréal, Québec, Canada

September 2014

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Ashkan Ebadi

Entitled: Impact of Funding on Scientific Output and Collaboration

and submitted in partial fulfillment of the requirements for the degree of

Ph.D. (Information Systems Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Michelle Nokken	Chair
Dr. Jorge E. Niosi	External Examiner
Dr. Ketra Schmitt	External to Program
Dr. Jamal Bentahar	Examiner
Dr. Chun Wang	Examiner
Dr. Andrea Schiffauerova	Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

Dean of Faculty

ABSTRACT

Impact of Funding on Scientific Output and Collaboration

Ashkan Ebadi, Ph.D.
Concordia University, 2014

This dissertation reports the results of a comprehensive quantitative analysis of the inter-relations among research funding, scientific output, and collaboration. The research employed various methods and methodologies (*i.e.* data and text mining, statistical analysis, social network analysis, bibliometrics, survey data analysis, and visualization techniques) to investigate the impact of influencing factors on researchers' performance, their amount of funding, and collaboration patterns. Moreover, a machine learning framework was suggested and validated for scientific evaluation of the researchers based on their productivity and level of funding. The Natural Sciences and Engineering Research Council of Canada (NSERC) was selected as the source of funding in this research since it is the main federal funding organization in Canada and almost all the Canadian researchers in natural sciences and engineering receive at least a basic research grant from NSERC. The required data on the scientific publications (*e.g.* co-authors, their affiliations, year of publication) was collected from Elsevier's Scopus. SCImago was selected for collecting the impact factor information of the journals in which the articles were published in as well as the annual citation counts of publications. The data was gathered and integrated for the time span of 1996 to 2010. The most significant contributions are: 1) the unique data extraction and gathering procedure that enhanced the accuracy of the target data, 2) the comprehensive triangulation technique which was employed in this research that included various methodologies and used new variables for assessing the inter-relations, 3) the proposed machine learning framework for classifying researchers and predicting their productivity and level of funding.

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Andrea Schiffauerova, for her continuous support, valuable guidance, and consistent encouragement throughout my research. This work was possible only because of your unconditional support, understanding, and immense knowledge. Thank you for providing me with the opportunity to learn.

I would like to thank my committee members for the brilliant comments and insightful suggestions. Also thank you to the faculty members of CIISE who helped me whenever I approached them. A special appreciation goes to Prof. Rachida Dssouli, Professor and Director of CIISE department, and Dr. Jamal Bentahar for their support. I am grateful to Mr. Carl St-Pierre for his valuable suggestions in statistical data analysis.

A special acknowledgement goes to all the friends, former and current members of the research group, you were amazing in too many ways. I enjoyed the friendly environment of the group in the past four years, thank you so much. This last word of acknowledgment I have saved for my wonderful family for their endless support, encouragement, and love.

CONTRIBUTION OF AUTHORS

This document is a manuscript-based thesis. The *results* section contains the results of the nine papers that were prepared and written in this research. All the nine papers were prepared under the supervision of Dr. Andrea Schiffauerova. Two other previously published journal articles are also listed as appendices.

Contents

List of Figures

List of Tables

1. INTRODUCTION

The link between the publicly funded research and scientific and knowledge-based systems is very important. Every year, governments spend large amounts of money on research mainly through universities and research institutes in expectation to improve the scientific potential of the country. It is thus essential to define good indicators for evaluating the research impact on the society, as well as to have effective measures in hand for making the best selection among the research groups competing for grants. Therefore, procedure of evaluating a research needs a group of indicators to create as precise a picture as possible of the various involved aspects in order to assess the performance of a researcher or a research group (King, 1987).

Scientometrics as a quantitative approach towards scientific development is not new. Alphonse de Candolle (1873), in an early study, highlighted the role of scientific societies in scientific strength of nations and tried to find effective factors for a nation's scientific success. Beginning with the qualitative methods (*i.e.* peer review) for the purpose of research evaluation (King, 1987), scientists tended gradually to more quantitative indicators. Lotka (1926), in his study on productivity of chemistry researchers differentiated from scientometrics a new stream called ***bibliometrics***. Since then, bibliometrics (a quantitative method) has been highly used in scientific research evaluations by applying the statistical and mathematical methods to books, articles and any other media of communication (Pritchard, 1969). However, there are still some doubts about the validity of bibliometric indicators to act as a single measure of scientific development (Glenisson, *et al.*, 2005). Several scientometrists tried to improve the performance of the traditional quantitative approaches and increase their evaluating power by introducing more complicated indicators or even new techniques (van Leeuwen, *et al.*, 2003).

It has been more than forty years that scientists tried to quantify the link between science and technology by collecting statistical data on scientific development (van Raan, 2005a). The use of large-scale bibliometric evaluation has originated in the USA (Hicks, *et al.*, 2004). In early 60s, the Organization for Economic Cooperation and Development (OECD) proposed a *standard practice for surveys of experimental research*

and development named *Frascati Manual* which took advantage of some standard measures enabling the government to collect information on research investment (van Raan, 2005a).

Funding has been acknowledged in the literature to be the main determinant of scientific development (*e.g.* Martin, 2003) and it is viewed as an important factor that has a significant effect on the scientific output since it provides a better access to the research resources (Lee & Bozeman, 2005). Despite all the efforts directed towards studying the outcomes of the funded research, with employing various methodologies and methods such as statistical approach, a variety of indicators, interviews, data pattern discovering, *etc.*, our knowledge about the subject is still limited (Godin, 2002). Moreover, most of the efforts were devoted to the study of the innovation process and not the results that will be gained (Cozzens, 2002). Although the studies have pointed out lots of benefits stemming from federally funded research, there are still many gaps in the evidence as a result of a variety of study fields and input sources (Salter & Martin, 2001). To justify the relation between costs of research and benefits that are gained more concrete and accurate evaluating mechanisms are required.

Apart from measuring the impact of funding on scientific output, several studies have examined its impact on the development of research cooperation and scientific collaboration both locally and internationally (*e.g.* Luukkonen, *et al.*, 1992; Leydesdorff & Wagner, 2009; Grossman, 2002). It was first after the World War II when the collaborations became tighter among researchers (Beaver, 1984). Despite the differences in defining the research collaboration, the importance of the collaborative research is now acknowledged in scientific communities (Wray, 2006), where higher financial investment can change the structure of research groups by increasing the collaboration among the scientists. However, the intensity of the impact of funding on scientific collaboration varies in different disciplines (Heffner, 1981). In addition to the efforts in measuring the effect of funding on the rate of collaboration, few studies analyzed the patterns of collaboration by creating the networks of the co-operation among researchers (*e.g.* Grossman, 2002; Hou, *et al.*, 2008). For this purpose, co-authorship analysis has been particularly recognized by some studies (*e.g.* Glanzel, 2001; Savanur & Srikanth, 2010) as being the most common tool in investigating the co-authorship relations and the quantitative patterns in scientific collaboration.

In order to understand the key elements that influence the link between allocated grants, scientific development and the structure of scientific collaboration in Canada, we need to have a better understanding of scientists' outputs and scientific outcomes generated per invested monetary unit. Moreover it is also necessary to shed some light on the collaboration patterns existing among the scientists and on the cooperation networks, within which the scientific output is generated. The aim of this thesis is thus to investigate the impact of the funded research and collaboration networks on scientific outcome in Canadian natural sciences and engineering. We decided to focus on natural sciences and engineering, because these disciplines can strongly enhance the economy, the society and the environment by means of technological innovation and discovery that can help a country to increase its scientific and technological capabilities. The methodology will involve a comprehensive approach including several methods and tools, *i.e.* bibliometrics, visualization techniques, statistical approaches, social network analysis, data and text mining, and survey data analysis.

The remainder of the document is organized as follows. The following chapter presents the background of the subject and reviews the respective literature. In Chapter 3, the objectives of the research are stated while Chapter 4 discusses data and methodologies used for the analysis. Chapter 5 reports the results in nine separate sections where each section presents a manuscript that was produced as the result of this research. Chapter 6 concludes, and Chapter 7 suggests some directions for the future research.

2. LITERATURE REVIEW

As technology and science develop, the competition among the countries increases. The research domains are being expanded with an expectation of higher quality and greater impact on the society. Since the governments of the developed countries devote a considerable part of the budget to research each year, it is understandable that they want to be able to evaluate the resulting outcome and research progress and to revise the allocation strategy accordingly (if required) (Gauthier, 1998).

In this section the relevant literature are reviewed to introduce the main topic of the thesis. It is divided into two main sections, highlighting different aspects of the main research topic. The first section discusses general methodologies which could be employed in the evaluation of scientific output. In the second section, the literature which addresses the funding concept and its impact on the scientific development and on the formation of scientific collaboration is reviewed in detail. Moreover, the research investigating the effect of collaboration on scientific productivity is analyzed. Descriptive and evaluative methods of measuring the science are compared and the most important factors which influence the evaluation of research performance are introduced. Finally, the research gaps are presented.

2.1 General Methodologies of Scientific Evaluation

After the Second World War (WWII) several industrialized countries started to devote more financial resources to research and development (R&D). Due to the large amount of investment, they decided to collect statistical data about R&D activities. Since then various methodologies have been employed to analyze research activities (Luwel, 2005). To evaluate how effective a research is, we have to measure the knowledge that it has produced! The challenge here is that the knowledge is intangible and it cannot thus be measured directly. Instead, we can only trace the evidence that such knowledge was generated through the scientific articles published in journals, presentations at conferences, patents registered with patent offices, *etc.* Luukkonen-Grunow (1987) and Averch (1990) categorized the main research evaluation methods into three categories; peer review, non-quantitative case study and quantitative methods. This is however a very general classification. In another study, Martin and Tang (2007) proposed a new classification of the methodologies with regard to the benefits of the funded research. In this section general research evaluation methodologies

that were mostly used in the scientific literature and are in line with the theme of this research are introduced and their advantages and weaknesses are discussed.

2.1.1 Peer Review

Peer review is one of the pioneer techniques and the most widely used method for research evaluation (King, 1987). It has been applied for a long time in different countries as a qualitative approach for evaluating the researchers' performance (Hicks, *et al.*, 2004). Although it is a fast and relatively low cost method, accuracy and quality of the peer review highly depends on the experts that are selected and also on the procedure and the criteria that are considered for the evaluation. King (1987) has mentioned the following limitations of peer review:

- Due to the preferences of peers, it is sometimes very difficult to find experts for some scientific areas.
- Expert review is useless for rearranging the scientific activities.
- More fame will result in getting higher funds.
- Reviewers may have different ideas about the research area.
- For the newer specialties, there would be no general agreement among the reviewers.
- Administrative costs and scientists' time which should be allocated to the peer review process is high.

Despite the aforesaid disadvantages, the great advantage of a peer review technique is that the impact of research could be assessed quite easily and accurately (Allen, *et al.*, 2009). For this important reason it has still remained as one of the most popular techniques in science evaluation, and is normally applied as a primary tool covering a wide range of methods. However, it is hard to find experts who are absolutely neutral (Arnold & Balaza, 1998) and the results could thus be easily influenced by subjective and personal views, and political and social external pressures. Hence it cannot be reliable enough as a single indicator, and the current trend is to combine the expert review with quantitative methods (Hicks, *et al.*, 2004) to achieve more accurate and fair evaluation.

2.1.2 Case Studies

In case studies, an evaluator selects a number of particular situations to study in order to understand the relations within the selected environment. This method narrates an event or a phenomenon using the in-hand data to support the findings. As an example, it can be employed to investigate how innovation occurs. Case studies are largely in use in R&D programs for evaluating the functional relationships. Particularly, case studies are suitable for identifying the interventions. They can help evaluators to gradually form the model and provide a narrative for the quantitative findings. Normally, they are used as a basis for more structured approaches (Arnold & Balaza, 1998) and can be useful to simplify the research and make it more understandable for non-scientific community (Ruegg, 2007). They are more performed in social and life sciences where explanatory case studies are used to find the underlying principles. Moreover, case studies and qualitative research are two completely different concepts where case studies can contain a combination of quantitative and qualitative approaches (Yin, 2009).

One of the main limitations of case studies is that since it is a narrative of the subject, therefore, it is less convincing in comparison with *e.g.* complex statistical techniques. In addition, the results of case studies could be inconsistent. However, they can be good information sources for the decision makers providing them with illustrative examples (Ruegg, 2007). Campbell *et al.* (2009) and Albrecht (2009) are two case studies which were done in the field of scientific evaluation and are discussed in the following sections.

2.1.3 Surveys

Surveys are useful tools for evaluating the progress of a program or statistically describing a program (Ruegg, 2007). They are based on the questionnaires that can be both quantitative and qualitative. The collected data can be employed for statistical analysis by testing the hypotheses. In addition, case studies or interviews can be used to validate the result of surveys (Arnold & Balaza, 1998). When the clean and reliable data is not available for a precise analysis surveys can be applied to generate the required data. They can be used in R&D evaluation to study the effects of a program or to act as a supplementary source of information (Ruegg, 2007). However, since surveys are highly dependent on the respondents' knowledge they could produce biased results (Martin & Tang, 2007).

Moreover, they usually suffer from the low response rate of the respondents which can highly affect the reliability of the results (Ruegg, 2007).

2.1.4 Econometric Studies

Econometrics has been widely used for studying the importance of research and development (R&D) and innovation (Loof & Heshmati, 2005). It mainly relies on statistical techniques (*e.g.* regressions) being applied on various databases. Since econometric studies mainly consider simplified assumptions for creating the model (Salter & Martin, 2001), they can be used as a good tool for testing the findings and results obtained through other methods and indicators (Arnold & Balaza, 1998) and for creating a general picture of the subject.

Various econometric approaches were used in the literature to measure the productivity of scientific systems. Input-output ratios were employed as simple measures of productivity. These methods are mainly employed as first order approximation. Farrell (1957) developed efficiency analysis which calculates a firm efficiency based on multiple given inputs and multiple taken outputs. He defined firm's efficiency as its success to produce as large as possible outputs given a set of inputs. However, when it comes to the scientific production and the analysis of science and technology systems it becomes complicated. In the scientific production the relations between inputs and outputs are uncertain and non-linear. Therefore, it is not a simple multi-input multi-output analysis, and external effects and internal relations should be considered as well (Bonaccorsi & Daraio, 2005).

2.1.5 Cost-Benefit Analysis

Cost-benefit approaches calculate returns from investment while considering total expenditure and the whole benefit that can be gained. They have been largely used by governments to evaluate their defined projects and to asses the R&D investment decisions (Ruegg, 2007). The problem is that both costs and benefits of research are practically very difficult to calculate especially if they are indirect (Lukkonen-Grunow, 1987).

Most of the cost-benefit analyses performed in the field of scientific performance evaluation are related to the health research. For performing such an analysis, the first move is to calculate the economic values of research to the society which is a very complex step.

According to the values, the appropriate level of investment is then decided. Such an analysis however usually encounters with a lot of potential obstacles. The main problem remains the inability to identify the exact values for input and output of the research. Moreover, it is rarely possible to appropriately detect the attributed economic impact (Buxton, *et al.*, 2004).

2.1.6 Bibliometrics

The need for accurate analysis of the science and technology policies is now obvious for governments (Okubo, 1997). One of the principal goals of evaluating the science by quantitative methods is to serve as an information tool for decision making (Gauthier, 1998). Rapid growth of information and the need for analyzing the useful information extracted from scientific publication databases developed into a new scientific discipline (van Raan, 1988). Bibliometrics, which could be applied to various other applications as well, is one of the quantitative methods most commonly used for the scientific evaluation and strategic decision making. Through bibliometrics we are looking for an overall picture of scientific output. One of the reasons that this method is increasingly being used for the evaluation purposes is that most of the available databases are suitable for applying bibliometric indicators (Lukkonen-Grunow, 1987).

With an aim to increase their evaluating power, more and more complicated bibliometric indicators have been developed (van Leeuwen, *et al.*, 2003). A wide range of bibliometric indicators have been used for assessing the scientific value of the research impact. Bibliometric indicators are specifically suitable for comparing large-scale patterns (Arnold & Balaza, 1998). Several studies have categorized bibliometric indicators. Rehn and Kronman (2008) in their *Bibliometric Handbook* divided the aforesaid indicators into three main categories as follows:

- Basic bibliometric indicators
- Advanced indicators
- Next generation indicators

Basic indicators are very simple ones, as they do not normally give an accurate picture of the studied area by themselves. Number of publications and citations during a particular time

period are two examples of very basic indicators. Due to the need for more exact and informative indicators for evaluating science, these indicators have been gradually more refined. Number of publications and citations per researcher, citations per publication, number of publications in high-impact journals, ISI (Institute for Scientific Information) journal Impact Factor, and h-index are some examples of improved basic indicators.

Advanced indicators take three important issues into account: publication year (due to the fact that older articles can be more cited), document type and research area (Rehn & Kronman, 2008). Moreover, there is always a normalization procedure needed for the advanced indicators. Two examples of these indicators are: field normalized citation score and top 5% (shows the number of publications related to a unit that belongs to the top 5% most cited publications, in the same year, subject and document type).

Various researchers and groups are now working on developing new indicators. As a work on next generation indicators, *Karolinska Intitutet* defined a project aiming to improve the current indicators. For this purpose, they focused on two separate categories, new subject classification and new statistical methods. Moreover, scientific indicators are becoming more and more developed reflecting the revolutionary changes in the web and web-related progress. The core citation-based impact indicators are still being used in studies, but they are supported now by some complementary techniques. An important factor which has played a role in changing bibliometrics is the availability of new information sources such as web pages and digital library usage statistics. To improve quality of the results, the current focus is on more precise data cleaning, on developing metrics for new tasks and on using bibliometrics to a wider range of problems (Thelwall, 2007).

Despite the wide range of applications, bibliometric indicators are faced with several problems in quantitative study of scientific activities. The main problem is the choice or the creation of a database. It is really important to have an integrated database that best suited to the needs of the particular research (Okubo, 1997). Apart from the database, citation itself is a complicated issue which makes the analysis that is based on it difficult to interpret. Although number of citations can be a good measure of the overall impact of an article, it cannot be a good factor of the article's quality (Seglen, 1992) due to various problems such as negative citations and self-citation.

2.1.6.1 Bibliometric Modern Indexes

Apart from the standard bibliometric indicators, *e.g.* number of publications, new indexes have been developed recently trying to better evaluate researchers' performance. The first modern measure of this category is h-index, introduced in 2005 by J. Hirsch. It relates an individual's published articles to the number of received citations. In order to calculate h-index, first the publications of a scientist are sorted based on the number of citations received. Then, h-index is calculated as the highest rank in a way that the first *h* papers received each at least *h* citations (Hirsch, 2005).

Later on and following the introduction of h-index, other scientists modified it and introduced new measures. The g-index (Egghe, 2006) aims to quantify the productivity of scientists according to their publication record. It is calculated as the highest number *g* of publications that together receives g^2 or more citations. Therefore, it is obvious by the definition that $g \geq h$. Jin (2006) realized that h-index does not consider the exact number of citations of papers included into the h-core[1]. He defined a new indicator and named it a-index. From this point of view, a-index is somewhat similar to g-index. A-index is calculated as the average number of citations received by the papers that are in h-core ('a' in a-index stands for average). In the same time, Kosmulski (2006) was working on another h-type indicator while trying to solve the problem of sorted long list of publications, which for a given scientist may require a time consuming calculation. He proposed a new scientific impact index that is called $h^{(2)}$*-index* or *Kosmulski-index*. With a list of papers in decreasing order of citations, $r=h^{(2)}$ is calculated as the highest rank that all the publications on ranks $1 \ldots h^{(2)}$ have at least $(h^{(2)})^2$ citations and the author is then said to have Kosmulski's index $h^{(2)}$. For example, if $h^{(2)}$-index for a given author is 5 then at least 5 of his papers have been cited at least 25 times each. Sidiropoulos *et al.* (2007) questioned h-index since scientists do not publish the same number of articles and, therefore, h-index could not be a fair measure. They normalized h-index and introduced h_{nom}. Egghe and Rousseau (2008) introduced citation-weighted h-index, a new h-index that is responsive to performance changes. This indicator is also called h_w-index. Zhang (2009) presented another h-index based indicator called e-index. It is a complement indicator to h-index and is very useful for evaluating the

[1] h-core is the total number of items (*e.g.* papers) that contribute to calculate h-index.

output of highly cited scientists or to compare research groups of identical h-index. In another study, Alonso *et al.* (2009) combined h-index and g-index to keep the advantages of the both measures and introduced a new indicator named hg-index calculated as the geometric mean of *h* and *g* indices of a scientist. Finally, Prathap (2011) proposed p-index, an indicator which is sensitive to performance and paper quality. P-index reflects to the scientific activity by considering the total number of citations along with the quality of the publications by taking the mean citation rate (citation per paper) into account.

2.1.7 Informetrics, Scientometrics & Webometrics

The term *scientometrics* was first introduced by Nalimov and Mulchenko in 1969 (Nalimov & Mulchenko, 1969). However, the term became more well-known upon the foundation of the journal '*Scientometrics'* by Tibur Braun in 1978. Scientometrics focuses on the scientific literature to quantify the aspects of science (Tague-Sutcliffe, 1992). In a more precise definition, Vinkler (2010) stated that scientometrics is not only the study of a scientific discipline but also study of people, groups, matters and phenomena in science and the relations among them.

It is really hard to specify distinct borders for scientometrics and bibliometrics. Tague-Sutcliffe (1992) believes that the mentioned fields overlap since they both focus on the quantitative study of publications. However, the focal points of these fields are a bit different. Bibliometrics concentrates on the scientific literature while scientometrics has a wider range of focus covering researchers' activities, organizations' policy, national economy *etc.* (Hood & Wilson, 1999).

Informetrics is a newer and more general term in comparison with bibliometrics and scientometrics which was first introduced by Nacke in 1979. Egghe and Rousseau (1990) highlighted in their book of "*Informetrics: Quantitative Methods in Library, Documentation and Information Science*" that Informetrics focuses mainly on the quantitative study of information. Moreover, according to Ingwersen and Christensen (1997), informetrics can even contain non-scholarly communities and the only requirement for the term is the production of information and its usage. In other words, informetrics is studying all sorts of available data and information (of any form) quantitatively to generate and distribute new information.

With an increase in use of the World Wide Web (WWW), new metrics have been recently created. *Netometrics* was introduced by Bossy in 1995 and is involved with the measurement of scientific interactions in the internet (Bossy, 1995). Two years later in 1997, *webometrics* was created by Almind and Ingwersen (1997) defined as the study of the network-based communications by means of informetrics (Almind & Ingwersen, 1997). Figure 1 shows the borders of the mentioned metrics.

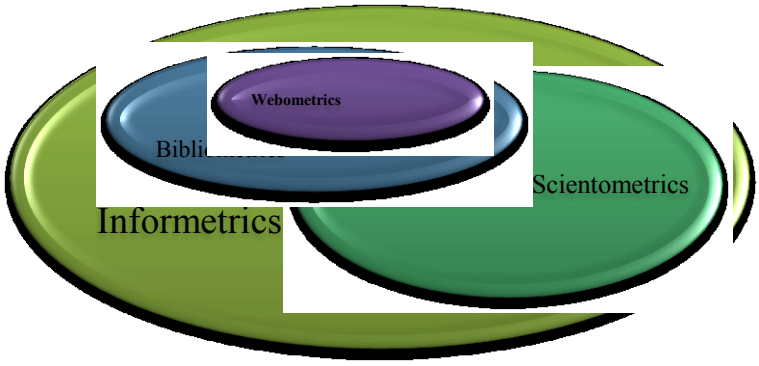


Figure 1. Borders of the new metrics (Björneborn & Ingwersen, 2004)

2.1.7.1 Main Informetrics Laws

Informetrics research is based on three important laws. The first one, Lotka's law, was introduced by Lotka (1926) and is related to the productivity of scientists. He named his discovery '*the frequency distribution of scientific productivity*' and later on was labeled Lotka's law by Zipf (1949). Lotka (1926) focused on chemistry and physics disciplines to study the number of contribution by different authors. When he plotted the number of publications over the number of authors the result was a Pareto-like distribution. Therefore, it can be stated that the number of authors with a specific number of publications is approximately equal to the inverse square of that number multiplied by the number of scientists who have just one paper (Wilson, 1999). For this reason it is also called the inverse square law. According to Lotka's law, a small number of scientists are publishing most of the scientific papers and the weights of publications are not divided evenly (Bookstein, 1980). Some scientists, *e.g.* de Solla Price (1976), made arguments on Lotka's law stating that the quantity of the scientific publications is not only related to the author's productivity but it is also affected by the span of time that a scientist is publishing actively (de Solla Price, 1976).

The second law is Bradford's law. He studied the distribution of a specific-subject related literature over journals. He declared that papers on a specific subject are normally published in a few related journals. However, he knew that this is not the real situation since the important publications in a subject are just a fraction of the whole literature that is being published in an increasing number of journals every year. Therefore, Bradford questioned subject related indexing of the literature and proposed source related indexing instead. Finally, he sorted journals in decreasing order of productivity and then plotted articles over journals and found that the distribution is Pareto-like. He split the journals into different zones of equal articles and found that the number of articles in each zone will increase exponentially (Bradford, 1934). The last law introduced by Zipf in 1935. He analyzed the words that are in the body of a specific scientific paper and ranked them based on their frequency. Zipf concluded that the result of ranks multiplied by the number of occurrences is constant ($r*f=c$) (Zipf, 1935).

2.1.8 Data Mining and Text Mining

Data mining is the process of extracting informative patterns from large databases. Although it is a newcomer as a scientific discipline, it is now employed in many fields such as statistics, information retrieval, machine learning, *etc.* (Hand, *et al.*, 2001). From the perspective of learning procedure which is employed, data mining approaches can be divided into supervised and unsupervised methods. In supervised methods (*e.g.* classification), learning is based on the training data which are accompanied by labels to define the class of the observations. However, in unsupervised methods (*e.g.* clustering), other cues such as Euclidean metric are applied on the input data to detect the classes or clusters in data (Han, 2006). Supervised methods are highly dependent on the training sets. On the other hand, unsupervised approaches try to fetch the patterns directly from the data.

Text mining is a special branch of data mining which is a suitable method for retrieving information out of scientific texts or text databases. Text Mining has originated from Information Retrieval (IR) discipline which mainly deals with storage, representation and access to information (Baeza-Yates & Ribeiro-Neto, 1999). Text mining methods have been mainly used for classification of the text documents so far. When dealing with extra large databases it is very helpful to categorize the available text. Clustering is one the

unsupervised methods which is known as one of the core data and text mining approaches. In this technique a similarity measure is defined and calculated for the individual items to categorize them into different groups. For this purpose, each document is introduced as a vector of words where the words frequencies are calculated. Then, clustering is done based on the vectors and the similarity measure (Leopold, *et al.*, 2005). Some of the supervised data mining techniques that were employed for classification purposes are as follows:

- Naïve Bayes
- K-Nearest Neighbors
- Support Vector Machines (SVM)
- Decision Trees

Naïve Bayes methods assume that the words of a document are generated through a probabilistic mechanism. By using the *Bayesian Formula* and a training set, documents are classified into different categories based on the similarities of their words to the words of that specific class (Dumais, *et al.*, 1998). In k-Nearest Neighbor, a similarity measure is used to select *k* training documents that are most similar to the test document. It is a non-parametric method and its good performance in practice have been approved (Joachims, 1998). Support Vector Machine (SVM) is an efficient and accurate technique for text classification tasks (Joachims, 1998; Dumais, *et al.*, 1998; Leopold & Kindermann, 2002). It uses a pre-classified training set to learn a decision boundary and then based on the learned boundaries it categorizes the input vectors. Learning in SVM is independent of the dimensionality of the feature space. This attribute enables SVM to be a very good approach especially for classification of texts (Leopold, *et al.*, 2005). Decision Tree (DT), as a standard tool in data mining, is a set of predefined rules which are employed sequentially for classification tasks (Mitchell, 1997). The main disadvantage of decision trees for text mining tasks is that the final decisions will be taken based on relatively few terms and conditions (Leopold, *et al.*, 2005).

Another application of text mining techniques is keyword detection. For this purpose the database is searched to identify important keywords and concepts. Moreover, text mining can be used to uncover the hidden relationships in the textual data (Ruegg, 2007). As an example, Roberts *et al.* (2005) used text mining to study the research relationships among

two units of the National Science Foundation (NSF). They detected the areas that were of mutual interests of the examined units. Co-word analysis is one of the other text mining techniques that is highly in use for scientific evaluation. This method is mainly employed as a tool for mapping different scientific fields. It is used for analyzing all that is inside a paper to find its relation to other publications. These kinds of indicators aim to form a structure for science. The result of co-word analysis is highly dependent on the variables and assumptions which have been considered for the model (Arnold & Balaza, 1998).

One of the main limitations of data and text mining techniques is that they require large amount of resources that make them relatively expensive. However, the availability of extensive digital text databases and faster computing systems has made these methods more attractive to the scientists recently (Ruegg, 2007).

2.1.9 Visualization Techniques and Maps of Science

Mapping techniques are two or three dimensional graphical representation of the structure of a scientific field. For this purpose, items that are related to each other are positioned in each other's vicinity. One of the main measures in building scientific maps is that two elements that are seen together in a same document could be identified as being closely related to be positioned in a map. Different elements then can be used to generate the scientific maps such as paper abstract, author(s) and cited references (Noyons, 2005).

Several studies have been done to generate maps of science. Van Raan (1996) performed a study to evaluate research performance using advanced bibliometric methods. He concluded that when we are aiming to map the socio-economical state of society, it would be essential to monitor both science and technology developments and those progresses which could be crucial in the near future. In another study, Leydesdorff and Rafols (2012) used Web-of-Science data to study the aggregate citation relations among 9,162 scientific journals and produced a global journal map. Porter and Youtie (2009) studied the position of nanotechnology field in the map of science. They used Science Citation Index (SCI) database and analyzed citation and publication data to explore the nature of research in nanoscience and the relationships among nanotechnology and other scientific fields. Boyack and Borner (2003) also used visualization approach to evaluate the impact of funding on

publications and found a positive relation between the rate of publication and funding amount.

2.1.10 Social Network Analysis

Social Network Analysis (SNA) investigates the structure of relationships among individual actors (representing as nodes) in a network aiming to reveal hidden and important connections (Ehrlich & Carboni, 2005). It has various applications in different scientific fields. In scientific evaluations, it can be used to analyze the links among researchers (or groups of researchers, organizations, *etc.*) and their development through visual mapping and measurement of the relationships. It is a useful approach for evaluating the impact of an R&D program or to study the important factors in forming scientific collaborations. One of the limitations of this method is that the generated network can be time limited and it might be required to regenerate the map to see the changes in the network (Ruegg, 2007).

SNA is highly related to the graph theory but it has its own terminology. However, the terminology varies across this scientific field may be due to its interdisciplinary nature (Freeman, 2004). A graph consists of a set of vertices (nodes, actors) and a set of lines (links, arcs, edges) between pairs of vertices. The degree is defined for each vertex indicating the number of lines that are incident with it. We call two vertices adjacent if they are connected by a line (de Nooy, *et al.*, 2005).

A large variety of metrics are used in performing social network analysis. One of the key measures is closeness centrality that highlights the importance of a vertex within a network. It is defined as the inverse of the sum of distances to all other vertices (Freeman, 1979). The other metric is the clustering coefficient that measures to what extent vertices tend to cluster together in a given graph (Holland & Leinhardt, 1971).

Actors (vertices) can play some special roles based on their position and characteristics in the network. In general, gatekeepers control access to something (*e.g.* information, knowledge, *etc.*) by making connections between two or more separate clusters (Gould & Fernandez, 1989). In the network analysis, this role can be detected by the betweenness centrality measure that is defined for each node as the proportion of all the shortest paths between the other nodes that contain that node (de Nooy, *et al.*, 2005). Other important actors are star scientists. "*Star scientist*" as a term was first introduced by Zucker and Darby

(1995) addressing an outstanding researcher with an excellent research productivity in terms of scientific and innovative activities. Therefore, star scientists are researchers with significantly higher productivity in comparison with their colleagues or other scientists. Degree centrality measure is used for this purpose to indicate important actors in the network and is defined for each node as the number of nodes that are directly connected to it (He, *et al.*, 2009).

2.2 Funding

Funding has been acknowledged in many articles to be the main determinant of research productivity (*e.g.* Martin, 2003). Having the literature reviewed, it is observed that the level of research funding is the most crucial factor for improving the research productivity. However, the approach towards the research funding varies across the countries. Different procedures are being followed worldwide for funding allocation. Some of them are based on the performance while others are based on the educational size. For example, UK is following a kind of performance-based approach for research funding. Performance-based evaluations, like other evaluation methods, have advantages and disadvantages. They enhance efficiency in short-run and create a better accountability. Moreover, they can be used for relating research to policy (Guena & Martin, 2003). However, the main problem of such evaluation is that getting reliable information is highly expensive (Bourke & Martin, 1992). In addition, if one can earn more from research rather than teaching by a performance-based funding system, professors will tend to the former and may cause publication inflation. On the other hand, educational size based funding systems have also some problems. These systems can give a very high power to the distributors of funds. In addition, it is hard to relate the number of the students to the scientific effort of that department. But, they are cheap and simple to operate. This makes such systems valuable (Geuna & Martin, 2003).

Considering the above, the essential question is: Do we get more benefits rather than costs by funding the research? Answering such a question is very hard since there will always be lack of input and output data and this makes the cost-benefit analysis difficult. In this section, I will first shed a light on the role of funding in scientific development and the benefits that will be gained through the publicly funded research. Then, the funding bodies

in Canada will be introduced briefly. In the next section, I will present a review of literature on the impact of funding on scientific productivity and collaboration among researchers while papers that have investigated the effect of collaboration on scientific productivity are also reviewed.

2.2.1 Role of Funding

About 100 years ago, the power and wealth of nations were measured by their amount of natural resources or the industrialization stage. Now it is knowledge which became a new worthy capital. In this respect, it is essential to strive to increase the production of the knowledge, which could be estimated by the research outcomes in terms of publication, scientific applications, and income (Oyo, *et al.*, 2008).

John H. Marburger, an American physicist who was the science advisor to President George W. Bush, in an editorial in Science in 2005, asked for a "*Science of Science Policy*" since investments in R&D have become more complex and challenging. He believed this could help policy makers to design more effective strategies (Leydesdorff & Wagner, 2009). To satisfy the need, the U.S. National Science Foundation formed an Interagency Task Group (ITG). ITG generated a road map in 2008 that addressed Marburger's recommendation in detail (ITG, 2008).

Funding can influence the size and efficiency of R&D sector and its productivity (Jacob & Lefgren, 2007). Different nations follow various research patterns and greatly differ in institutional and economic structures. In some countries (*e.g.* Sweden) more than 3% of Gross Domestic Product (GDP) is spent on research and development (R&D) while others spend less than 2%, *e.g.* UK (Leydesdorff & Wagner, 2009). In Canada, the ratio of Gross Expenditure on R&D (GERD) to GDP was about 2 in the last ten years which is not a good rate in comparison with other developed countries. Table 1 shows GERD/GDP ratio in Canada for the period of 2000-2009. Although the amount of the expenditure on R&D has increased during the past ten years, the GERD/GDP ratio remained almost constant or even decreasing due to the increase in the amount of GDP.

Table 1. GERD to GDP ratio in Canada, 2000-2009 (Statistics Canada, 2010a)

Year	GERD/GDP	Year	GERD/GDP
2000	1.91	**2005**	2.04
2001	2.09	**2006**	2.00
2002	2.04	**2007**	1.96
2003	2.04	**2008**	1.87
2004	2.07	**2009**	1.92

Furthermore, the composition of the budget which different countries are allocating to R&D varies. As a result, various allocation patterns are available world-wide to distribute the research budget among the universities, research institutes and others (Leydesdorff & Wagner, 2009). In Canada, R&D expenditures are divided into two major scientific fields which are natural sciences and engineering, and, social sciences and humanities. 90% of the total R&D expenditures are dedicated to the category of natural sciences and engineering (Statistics Canada, 2010b).

2.2.2 Benefits of the Publicly Funded Research

The areas in which research can generate benefits are wide. Linking the research and its impacts on the society could be a very challenging issue. Geisler (2004) provided a very practical flowchart of such a link which is shown in Figure 2.

Figure 2. From primary outputs to final outcomes of research (Geisler, 2004)

One of the reasons that scientists publish their work in the form of scientific papers is that in this way they can secure their priority in discoveries (De Bellis, 2009). As it can be seen in Figure 2, it takes some stages to move from the primary benefits of research to final outcomes. The more complex the outcome the more time it needs to be achieved. However,

the main reason that governments invest in research is for its final outcomes not just for the sake of the research (Hicks, *et al.*, 2004).

Martin *et al.* (1996) counted six major benefits that are obtained from the publicly funded research:

- Increasing the available knowledge
- Creating and improving scientific technologies
- Motivating social interaction through collaboration networks
- Training skillful graduates
- Creating new jobs and companies
- Increasing the problem-solving ability of the researchers

Although some of the above mentioned benefits are interrelated, it is good to separate them analytically. An example of the overlap can be the categories of "*Training skillful graduates*" and "*Increasing the problem-solving ability of the researchers*" that are interrelated but not exactly identical. These benefits can be obtained from any funded research regardless of the source of the funding that can be private or public (Salter & Martin, 2001).

2.2.3 Funding Bodies in Canada

There are three main funding bodies working in Canada:

- The Canadian Institute of Health Research (CIHR)
- The Natural Sciences and Engineering Research Council (NSERC)
- The Social Sciences and Humanities Research Council of Canada (SSHRC)

CIHR, established in 2000, is responsible for the support of health research in Canada. It is managed by the Prime Minister and the Governing Council. In fiscal year 2006-2007, CIHR invested over $832.7 million in research projects and personnel support, out of which $660.7 million was the amount of grants budget (CIHR, 2012).

NSERC, established in 1978, is a Canadian government agency that provides funds for research. NSERC reports to the Parliament through the Ministry of Industry. NSERC

supports about 23,000 university students as well as 11,000 university professors. Budget of NSERC for funding programs in 2005-2006 was $859 million (NSERC, 2012).

SSHRC was established in 1978 and is a subsidiary of Canadian federal government agency. It supports a wide range of research in social sciences and humanities. It offers several types of funds *e.g.* researcher-framed grants, research fellowship and student funding (SSHRC, 2013). About 19,000 universities and college faculties (53% of the total number of academic community of Canada) and 49,000 full-time graduate students (55% of the Canadian total) are working in social sciences and humanities fields. The budget of SSHRC for grants and scholarships in the year 2006-2007 was $306.2 million supporting around 5,200 researchers (Mitchell, 2006).

The public research funds are allocated by each of the mentioned granting bodies through a peer review evaluation of the research proposals on a project by project basis. Even though so great amounts of funding are involved, it has been suggested that the research outputs are not being systematically evaluated in Canada (T.A.C. Group, 2005).

2.3 Funding Impact on Scientific Output and Collaboration

This section critically reviews the papers that studied impact of funding on scientific output and collaboration. First, the impact of funding on scientific output is investigated and then its effect on scientific collaboration is assessed.

2.3.1 Funding Impact on Scientific Output

Investigating the impact of funding on the quality and quantity of the published research has attracted more attention of the scientometrists in comparison with analyzing funding effect on collaboration. It is easy to judge the productivity and the impact of the research of the Nobel laureates. However, for the rest of scientists one should have quantitative indicators in order to analyze and compare the scientific productivity of the researchers (Hirsch, 2005). The number of publications has been widely used in the literature as the quantity proxy of scientific activities. However, according to Okubo (1997) publication counts can be considered as a reliable measure just for large-scale data (*e.g.* macro-level, cross-countries level).

It is generally accepted in bibliometrics that the real or expected number of citations received by publications can be used as a good index of the mean impact at the aggregate level (Gingras, 1996). However, the citations have several drawbacks and thus are considered by some (*e.g.* Seglen, 1992) as a poor measure of quality. As an example, papers of famous scientists are more likely to be cited. One of the reasons is that they normally supervise a lot of students and they have different teams working on various projects. Apart from that, a low quality work may receive many negative citations, *i.e.* it is cited not due to its quality but due to an error in methodology or results (Okubo, 1997).

Evaluating the relation between research input (*e.g.* research funding) and research output (*e.g.* number of publications) has been a challenging issue for policy makers. A number of techniques (*e.g.* scientometrics, statistical analysis) can be used for this evaluation (King, 1987). It is generally assumed that funding has a positive effect on scientific development and number of scientific publications (Campbell, *et al.*, 2010; Boyack & Borner, 2003; McAllister & Narin, 1983; Godin, 2003; Campbell, *et al.*, 2009). Apart from the number of publications, the impact of funding on the quality of published papers has been also studied. In this section, first I discuss the studies that analyzed the impact of funding on the quantity and quality of the output. Then, the studies that have evaluated the impact of funding at the macro-level are investigated and, finally, I will specifically discuss the studies that have been done in Canada so far.

2.3.1.1 Funding Impact on Quantity and Quality of Scientific Output

In an early case study performed by McAllister and Narin (1983) for the National Institute of Health (NIH) the relation between NIH's funding and number of publications of the U.S. medical schools was investigated. Using bibliometric indicators they found a quite strong relationship between the funding and the number of papers published. Moreover, they found that the number of papers and their citations for each medical school are well related to the quality of the school. Their results partially indicated that funded research may be more cited than the unfunded ones. In a similar study Peritz (1990) analyzed the citation impact of funded and unfunded research in the field of economics. Using statistical analysis, he found that even if both funded and unfunded researches are published in a high-impact journal, the funded research will be more cited, which is in line with the findings of

McAllister and Narin (1983). Although he found a positive impact of funding on the quality of the output, some criticism related to his method was raised. He performed a significance test but he did not consider a random model. This approach has been criticized in the literature since if the sample is not randomly taken then the significance test may overstate the accuracy of the results.

A few studies investigated the effect of funding on the output of medical (health) schools (programs). Lewison and Dawson (1998) studied the funding effects on the outputs of biomedical research. They used journal impact factors as a quality measure with a small modification (applied a five-year citation period) to overcome the short-term influence problem which such indicators may have. They concluded that the number of authors per article and the number of funding bodies both have a great effect on the impact of research output. With the increase in the number of authors, an increase in multi-disciplinarity could be observed. This is considered to be a highly important factor for an increase in the impact of the research output. More specifically, they found that if the number of authors rises from one to six then the mean journal impact increases more than twice while the number of citations received is tripled. Jacob and Lefgren (2007) analyzed the effectiveness of government expenditures in R&D by investigating the impact of NIH funding on the quantity and quality of the papers of the funded researchers. Their database contained researchers who were funded by NIH in 1980-2000. They used OLS regression to perform the analysis. According to their results, NIH grants had a positive impact on the publication rate leading to about one additional publication over the next five years. This positive impact was higher for postdoctoral fellows. Cancer Association of South Africa (CANSA) conducted a research to evaluate the quality and quantity of the funded research during a 10-year period (1994-2003). In this study, Albrecht (2009) took advantage of bibliometrics and counted the number of peer-reviewed publications in PubMed database for each grantee which were also related to CANSA grants. Since CANSA grants were partial he could not create a benchmark for the cost of an average, peer-reviewed cancer research publication in South Africa. However, he found that the research was more focused on the areas of cancer biology and experimental treatment.

Arora and Gambardella (1998) analyzed the impact of the contractual funding on Italian academic researchers who work in the biotechnology field. They defined a scientific

research production function including various variables such as budget requested, budget granted, size of the group, age of the principal investigator (PI), and number of papers adjusted by quality. They found that although the average elasticity[2] of the output with respect to the funding is around 0.6, the most reputed research groups have elasticity close to 1. In addition, they realized more unequal distribution of funds may increase the output in the short term. However, they had some limitations in performing their analysis such as lack of micro-level data on funding levels and research output in various scientific fields. Carayol and Matt (2006) studied some important factors that affect quantity and quality of scientific production of the faculty members of Louis Pasteur University. Based on their funding variables, they concluded that the effect of private contractual funding is not significant. However, research output is positively influenced by the public contractual funding. But even in this case the respective coefficient is very small.

Payne and Siow (2003) analyzed the impact of federal funding on 74 research universities. Employing a regression analysis on a panel data set spanning from 1972 to 1998, they investigated the effects of funding on the articles published and patents issued by the researchers. Their results show a small positive impact of funding on the number of patents while the effect on the number of articles is relatively higher ($1 million leads to 11 more articles and 0.2 more patents). They could not find a significant impact of funding on the quality of the articles measured by number of citations per article. In an econometric evaluation of the impact of funding composition on agricultural productivity, Huffman and Evenson (2005) used annual data for 48 U.S. states from 1970 to 1999. They found a significant negative impact of the federal competitive grant funding on the productivity of public agricultural researchers. Gulbrandsen and Smeby (2005) studied the effect of industry funding on the performance of university professors in Norway. They used questionnaires to collect data from all tenured professors in Norway and employed logistic regression to perform the analysis and found a positive relation between the industry funding and researchers' performance.

In two recent studies, Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) studied the impact of public and private funding on the scientific production of the

[2] Elasticity measures how changing one variable affect other variables. Small changes can have large effects on elastic variables.

Canadian academics working in biotechnology and nanotechnology fields respectively. Beaudry and Clerk-Lamalice (2010) defined their regression model including structural network properties variables, universities dummy variables, and grant and contract amounts. They found no negative impact of contracts on the publication output of researchers working in biotechnology. However, a positive effect of funding and strong network position on the scientific output was observed. The regression model of Beaudry and Allaoui (2012) is very similar to the Beaudry and Clerk-Lamalice (2010) model. They added number of patents and age of the researchers variables to their model and assessed the impact of the considered variables on the scientific output of nanotechnology researchers. Although they found a positive effect of public funding on scientific publications, the effect of private funding on scientific output is inexistent. They have considered a narrow but highly multi-disciplinary field, nanotechnology, to compare the scientific production of universities. Moreover, their model cannot assess the effect of graduate students on scientific production.

In a different attempt to relate funding to the scientific development, researchers have recently focused on 3D mapping of the grants and publication data using the visualization tools such as VxInsight©. Boyack and Borner (2003) evaluated the influence of grants on publications. By using VxInsight map, they found a positive relation between the allocated funds and the publication rate in most of the cases. They have also included a 3D map combining the grant and publication data together in one picture trying to better visualize the impacts. They propose that although such resulting maps cannot replace human decision making, the researcher or government workers can use them to accelerate their understanding of large data sets and to facilitate the decision making procedure. This is a pioneer study in 3D visualizing of funding and scientific output data together in one map. However, they faced with some limitations such as lack of more accurate data and using larger amounts of data that might need more efficient and more complex data mining and clustering techniques. In the next section, the literature that analyzed the impact of funding on scientific output at the cross-countries level (macro-level) is reviewed.

2.3.1.2 Funding Impact on Scientific Output, Macro-level

Some researchers have studied the impact of financial investment on scientific production at cross-country level. Leydesdorff and Wagner (2009) analyzed the relation

between research macro-level investment and world share of publication. They employed the main science and technology (S&T) indicators of OECD (2008). They found a lot of differences among examined countries in terms of their efficiency in turning financial investment into scientific output. Apart from the efficiency issue, they found different schemes of funding in various countries. In an econometric study, Crespi and Geuna (2008) analyzed the important factors (especially the investment) that influence scientific productivity. They focused on the higher education in 14 OECD countries and used Thomson's ISI database to gather the publication and citation data for the period of 1981-2002. They mainly focused on the time lag structure of the output and the nature of the spillovers and concluded that investment had a significant impact. In a very brief study, Shapira and Wang (2010) investigated the impact of nanotechnology funding. They used Thomson Reuter's database for the period of August 2008 to July 2009 and used very basic bibliometric indicators to give a general picture of countries which are working in the nanotechnology field. They argued that as an impact of large investment that has been made, China is getting closer to the U.S. in terms of the number of publications but Chinese papers still have lower quality in comparison with the Americans and Europeans.

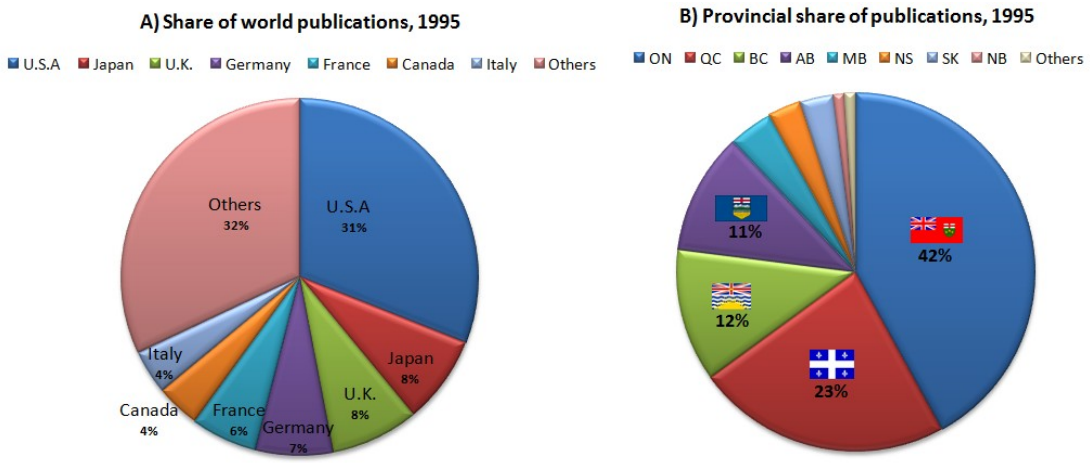
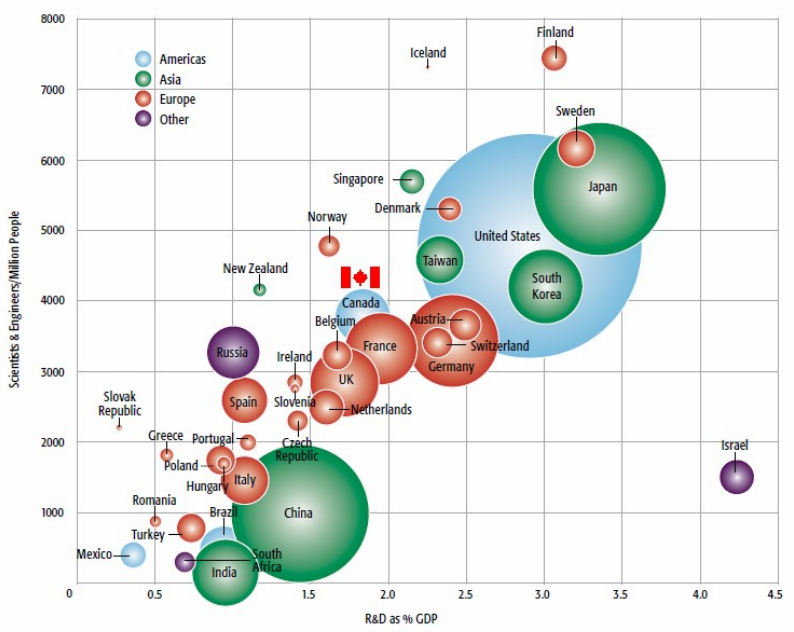


Figure 3. Canada scientific output, 1995
Source: ***Observatoire des sciences et des technologies (CIRST), March 1998***

To elaborate on the importance of cross-countries evaluation, a simple example follows. One of the primary measures that has been used for indicating the research output at the international level is the number of a country's scientific publication that can be used as a basic indicator to compare the scientific performance of different countries. In addition, by

combining the results with the amount of countries' investments their worldwide position can be identified. Figure 3 shows different aspects of scientific output in Canada (Gauthier, 1998). As shown in Figure 3-A, based on publication indicator Canada is among 7 leading countries. However, the share of USA is much greater than the others representing about one third of the whole. Moreover, Figure 3-B shows the share of different provinces of Canada in scientific development of the country in 1995 (Gauthier, 1998). It can be easily observed that Ontario, Quebec and British Colombia are the key players in producing scientific knowledge in the country.



*Circle size reflects the relative amount of annual R&D spending by the country noted.

Figure 4. World R&D Expenditures, 2010 (Grueber & Studt, 2010)

According to Figure 4, the level of Canada funding in 2010 was about 1.8% of Gross Domestic Product (GDP)[3]. Figure 3-A and Figure 4 are implicitly confirming the positive relation between the amount of investment on R&D by a country and the scientific output.

2.3.1.3 Funding Impact on Scientific Output, Case of Canada

The evaluation of research performance in Canada has started attracting the attention of the policy makers recently. In Canada, scientific articles have been recognized as the main output of researchers and universities (Godin, 2003) and bibliometrics has been used for

[3] GDP is the value of all final goods and services produced in a country within a given period.

scientific evaluation purposes. This section discusses the studies that investigated the role of funding and investment on the productivity of the researchers in Canada so far.

Gingras (1996) in a report to the Program Evaluation Committee of NSERC discussed the feasibility of bibliometric evaluation of the funded research. He focused on two grant selection committees (Mechanical Engineering and Evolution & Ecology) aiming to find whether the results of the indicators and data which come from bibliometrics can be used for answering the questions about funding allocation policies. He showed that it is applicable to use bibliometric indicators to investigate the relation between funds and scientific productivity since these measures give considerable information about such relations. Furthermore, his analysis indicated that it is feasible to apply evaluative bibliometric methods on the funded research at the disciplines or specialties level. Gingras (1996) employed simple bibliometric indicators for performing his analysis and did not perform any statistical analysis.

Following his study, a few Canadian researchers used bibliometrics for analyzing the funding impact. Godin (2003) in a bibliometric evaluation studied the impact of NSERC funding on the productivity and papers' quality of the supported researchers for the period of 1990-1999. He used Science Citation Index (SCI) database and analyzed the number of papers written by funded researchers over a 10-year time period to find NSERC proportion amount of contribution to the scientific development of Canada. For this purpose, he applied two indicators (ratio of papers of the funded researchers which were written in collaboration with others and the journal quality in which the funded researchers publish their papers). He found that researchers with higher amount of funding available are more productive. In addition, when the level of funding for a given researcher is above the median (high) his/her productivity is more strongly correlated with the amount of funding. However, the level of funding does not affect the researchers' journals quality. These results are based on simple bibliometric indicators hence, no strategy in regard with better allocation and distribution of grants can be set.

In a series of case studies, Campbell and his colleagues performed bibliometric evaluations on the impact of funding on scientific performance. Campbell *et al.* (2010) utilized bibliometrics as the performance measurement tool for evaluating the impact of the

research which was funded by the National Cancer Institute of Canada (NCIC). They worked on Thomson Reuters' Web of Science (WoS) database (which includes three databases: Science Citation Index (SCI) Expanded, the Social Sciences Citation Index and the Art & Humanities Citation Index) trying to cover all fields of science to get the respective statistics of NCIC's funded researchers. They calculated two main bibliometric indicators in this respect, *i.e.* number of papers and average of relative citation (ARC). Besides, some statistical tests were performed to check the differences between the scientific impacts of different entries. Their findings show a positive relation between the funds that have been provided by NCIC and the scientific performance. In other words, the ARC of NCIC-supported papers were of higher value than those of non-supported ones. In a conference presentation Campbell and Bertrand (2009) reviewed the results of a bibliometric measurement of research performance for the Canadian Forest Service (CFS). They used a quite wide range of bibliometric indicators to assess CFS internal, national and international position. They found that CFS has the most papers published in forestry in Canada and internationally it ranks as 3rd. Although there were some fluctuations in the impact of CFS publications during 1990-2002, the results showed that it has increased to above the world level during 2003-2006. Finally, they found that CFS ranked 3rd in number of collaborations within the top 50 world institutions network. In another similar research, Campbell *et al.* (2009) evaluated specifically the selection procedure of Genome Canada to see whether it allocates the funds to the right researchers. By means of Thomson Reuters' Web of Science (WoS) database and many bibliometric indicators (number of publications, ARC, number of cited papers, number of papers in highest impact journals, *etc.*), they claimed that the peer-review process of Genome Canada was successful in researchers selection. Moreover, the papers published by Genome-funded researchers have a significantly higher scientific impact than other genomics papers not just in Canada, but also all over the world. That means there was a positive relation between Genome-funded projects and the scientific performance of the supported researchers. Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) are the other two studies that were done in Canada and were discussed before.

2.3.1.4 Funding Impact on Scientific Output, Summary

For over sixty years, governments have funded researches (Godin & Doré, 2004) and have used various tools and techniques, quantitative and qualitative, to measure the scientific performance (Godin, 2002). This section summarizes the reviewed literature about the subject.

Table 2. Summary of the research on evaluating the impact of funding on scientific output

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
McAllister and Narin	1983	Bibliometric indicators	---	NIH funded researchers	- Strong relationship - High quality schools are more productive
Peritz	1990	Statistical analysis	Articles published in two British journals in 1978 and 1979 and their citations	Economics	Funded researchers are being more cited
Gingras	1996	Feasibility study Bibliometric indicators	1984-1993	NSERC funded research	Feasible to apply evaluative bibliometric methods on the funded research at the disciplines or specialties level
Arora and Gambardella	1998	Econometrics OLS regression	1989-1993	Biotechnology, bio-instrumentation Researchers sponsored by the Italian National Research Council (CNR)	More unequal distribution of funds may increase the output in the short term
Lewison and Dawson	1998	Statistical analysis	Research Outputs Database (ROD) 12,925 UK papers 1988-1994	Biomedical (gastroenterology)	Positive relation between the number of funding bodies and research output impact
Boyack and Borner	2002	Visualization 3D mapping	33,448 grants records 4,549 outputs 1975-2001	Behavioral and Social Science (BSR) program Output resulted from National Institute on Aging (NIA) grants	Positive relation between funding and output in most of their cases
Godin	2003	Bibliometrics	Science Citation Index (SCI) database 1990-1999	NSERC funded research	Positive relation between funding and productivity, but no impact on quality
Payne and Siow	2003	Regression analysis	Federal funding 1972-1998	74 research universities	Small positive effect on the number of patent and a larger positive effect on the number of articles No impact on research quality
Huffman and Evenson	2005	Econometrics	USDA 1970-1999	Agricultural research productivity	Negative impact of the federal competitive grant funding on the productivity of public agricultural researchers
Carayol and Matt	2006	Statistical analysis Linear regression OLS model	1993-2000	Faculty members of Louis Pasteur University (ULP)	No significant effect of contractual funding, except for the public one where the coefficient is a small positive number
Jacob and Lefgren	2007	OLS and regression	1980-2000	NIH funded researchers	Positive impact on the rate of publication
Crespi and Geuna	2008	Econometric	Thomson ISI database 1981-2002	Higher education in 14 OECD countries	Investment have a significant impact on the time lag structure of the output
Albrecht	2009	Bibliometric indicators	PubMed database 1994-2003	Cancer Association of South Africa (CANSA)	No conclusion in regard with impact of funding on output

Table 2. Continue.

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
Leydesdorff and Wagner	2009	Main S&T indicators of OECD (2008)	Thomson's Web of Science	Macro-level comparisons	At cross-country level, lots of differences are observed in the link between investment and world-share of publication
Campbell *et al.*	2010	Bibliometrics	Thomson Reuters' Web of Science (WoS) database	Researchers funded by National Cancer Institute of Canada (NCIC)	Positive relation between funds and output quality
Shapira and Wang	2010	Bibliometric indicators	Thomson ISI database 2008-2009	Cross-country evaluation	Positive impact of China's investment on output quantity, but no major effect on the quality
Beaudry and Clerk-lamalice	2010	Regression analysis	1985-2005 3,678 articles	Canadian biotechnology academics	No negative impact of contracts on publication Positive effect of strong network position and individual funding on scientific output
Beaudry and Allaoui	2012	Regression analysis	Articles and patents 1985-2005 3,724 articles 566 patents	Canadian nanotechnology researchers	Positive effect of public funding, no impact if private funding

As it can be seen in Table 2, the most dominant approach that has been used in analyzing impact of funding on the scientific productivity is the statistical analysis. Table 3 shows different determinant factors that have been considered in the literature in the respective statistical models. Regional share and scientific field variables have been the most attractive variables for the researchers while paper quality has been also considered as an important factor. From Table 2 and by counting the number of recent studies, it is clear that this issue is becoming more important and it is getting more attention of the researchers. Some reasons could be the increase in the number of the authors, the limited sources of funding available, and recent economic depressions that forced the policy makers to reevaluate their strategies and design them in a way that would stimulate scientific development more efficiently.

Among the studies that performed regression analysis just Arora and Gambardella (1998) used a control group of non-funded researchers. However, they have not used it to assess the net impact of funding on the scientific output, but they employed the information from non-funded units to estimate the grant selection and resource allocation equation. Since the other two studies that used the control group (Peritz, 1990; Campbell, *et al.*, 2010) performed descriptive analysis, the net impact of funding on scientific output is still vague. In other words, apart from the funding variables that directly affect the scientific productivity, other factors can also affect the output indirectly. For example, higher funding may influence scientific collaboration by forming more efficient groups that may lead to higher scientific productivity. Therefore, one may notice that the collaboration network structural variables

can be important factors in determining the productivity of the researchers. Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) have recently considered this issue in their study by adding two network structure variables.

Table 3. Statistical analyses and variables, impact of funding on scientific output

Article	Ctrl grp	Independent Variables*							Type of analysis	Various sources of funding	Different funding periods	Network structure vars
		Fund	Other produc-tivity shocks**	Prestige/ career/ Age	Scientific fields	Regional Share	Grp size	Paper quality				
Peritz (1990)	✓	✓			✓				Descriptive analysis	No	No	No
Arora and Gambardella (1998)	✓	✓		✓		✓	✓		OLS regression	No	No	No
Lewison and Dawson (1998)		✓				✓	✓	✓	Statistical analysis	No	No	No
Godin (2003)		✓			✓	✓		✓	Descriptive analysis	No	No	No
Payne and Siow (2003)		✓	✓					✓	Linear regression	No	No	No
Huffman and Evenson (2005)		✓				✓			Econometr-ics	Yes	No	No
Carayol and Matt (2006)		✓	✓	✓	✓				Statistical analysis OLS regression	No	No	No
Jacob and Lefgren (2007)		✓	✓		✓			✓	OLS regression	No	Yes	No
Crespi and Geuna (2008)		✓				✓		✓	Econometr-ics	No	Yes	No
Campbell *et al.* (2010)	✓	✓				✓		✓	Descriptive analysis	No	No	No
Beaudry and Clerk-lamalice (2010)		✓				✓			Negative binomial regression	Yes	No	Yes
Beaudry and Allaoui (2012)		✓		✓		✓			Negative binomial regression	Yes	No	Yes

* No of articles (or similar variables such authors/paper) has considered as the dependent variable
** *E.g.* The quality of institution, external shocks to schools, past publication pattern *etc.*

Apart from Crespi and Geuna (2008) that studied the time lag structure of the scientific output in 14 OECD countries, just Jacob and Lefgren (2007) considered a couple of funding independent variables reflecting different periods of funding (including pre-funding period). Considering such a factor can help to better analyze the impact of funding on scientific output by distinguishing the impact of each period. However, Jacob and Lefgren (2007) used OLS regression on a limited set of variables while Crespi and Guena (2008) used non-linear econometrics approach indicating the combined impact of the independent variables.

Another important point is that increasing the number of independent variables may augment the risk of having correlations among the variables. In order to calculate the net impact of funding on the scientific output more independent variables of different types are

required. To overcome the risk of having correlations among the included variables one way is to increase the size of the population. Apart from the limited number of variables in the mentioned studies, big population size has been also neglected by most of the researchers.

2.3.2 Funding Impact on Scientific Collaboration

Scientific activities know no borders. All researchers worldwide are working together in a global community to improve the level of knowledge. However, technological developments that are the applications of scientific knowledge differ from the supra-national nature of the scientific activities that highly rely on the prior knowledge (Subramanyam, 1983). Due to the nature of the modern science which has become more complex and inter-disciplinary, scientists may tend to collaborate more.

According to Katz and Martin (1997) scientific collaboration is defined as the process through which the researchers with a common goal work together to produce new scientific knowledge. Scientific collaboration has been studied in a vast number of different disciplines such as computer science, sociology, research policy, and philosophy (Sonnenwald, 2007). In addition, various types of collaboration have been mentioned in the literature including inter-firm collaboration, international collaboration, and academic collaboration (Subramanyam, 1983). This diversity in examining the scientific collaboration and its different types makes it more probable to find various methods, approaches and terminologies for this purpose in the literature (Sonnenwald, 2007).

Measuring the scientific collaboration is not easy. Although co-authorship and sub-authorship [4] have been both considered in the literature as indicators of scientific collaboration, only 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 as a measure is practical, invariant, verifiable, inexpensive (Subramanyam, 1983), and quantifiable (Katz & Martin, 1997). Through co-authorship researchers get access to an often informal network of scientists that may facilitate knowledge and skill diffusion (Tijssen, *et al.*, 1996; Tijssen, 2004). There are also some drawbacks in using co-authorship as an indicator of collaboration since collaboration

[4] Measured by the number of researchers/colleagues thanked in the acknowledgement section (Sonnenwald, 2007).

does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997).

Collaboration can generate large advantages for the society. Through the collaborative scientific activities, different skills and ideas are combined, and resources are thus used more efficiently. This can bring economies of scale in scientific activities and may avoid research duplication (Ubfal & Maffioli, 2011). In addition, collaboration trains the available skills that will result in the development of new expertise (Lee & Bozeman, 2005). However, there are some costs (*e.g.* finding right partners and research coordination) related to the scientific collaboration that may influence the optimal individual collaboration level *(*He, *et al.*, 2009). 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. However, Adams *et al.* (2005) evidenced that the scientific collaboration cost has declined in the last two decades, which might be explained by the lower travelling costs and improved communication technology.

Although governmental funding for knowledge creation and diffusion has a long history, its effects on scientific collaboration and formation of scientific networks is relatively new (Katz & Martin, 1997; Lee & Bozeman, 2005). The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006), where financial investment can change the structure of research groups and affect the collaboration among the scientists. However, there might be some conflicts between individual preferences and the society level goals. These conflicts may cause different optimal individual collaboration level from the optimal social one. Therefore, the efficient collaboration network will not be stable since the central actor(s)[5] bears a huge coordination cost that is not of his/her private interest. As a result, to evaluate the policies that affect the collaboration the relation between the individual incentives and social benefits should be considered as well (Ubfal & Maffioli, 2011).

[5] More central actors have higher degree or more connections and tend to have favored positions in the network.

A great advantage of public funding is that it enables researchers to cover the collaboration costs. Moreover, it allows the central actors to better internalize some of the required duties through the coordination (Ubfal & Maffioli, 2011). According to Porac *et al.* (2004) availability of funding may help the central scientist(s) to make a balance between the new knowledge creation and the management of the existing collaborative relationships in the network. On the other hand, one may notice that higher amount of funding is not always good. If the collaboration network is at its social optimum level then allocating more funding can affect the system negatively by adding more collaboration links (Ubfal & Maffioli, 2011). This scenario normally happens in developed countries and is not the case in developing countries (García de Fanelli & Estébanez, 2007).

There are only a few studies that specifically investigated the effects of funding on scientific collaboration. However, most of them are limited to the performance of universities or educational environments. Also, they have not considered a test group of non-funded researchers. This can help to determine the net impact of funding. In addition, to my knowledge no efforts have been done towards analyzing the funding impact on the structure and pattern of collaboration networks.

In an early study, Beaver and Rosen (1979) studied the effect of funding on the average number of authors per article as an indicator of the scientific collaboration in 24 scientific fields and found a positive relation between funding and the average number of authors per article. Two years later, Heffner (1981) collected 500 articles published in 28 journals in four scientific fields during 1974-1975 and analyzed the relation between funding and multiple authorship statistically. He found that with an increase in the financial support, size of the research teams has generally increased but the impact of funding was statistically significant just in chemistry and biology (two fields out of the four examined fields).

Using questionnaires for gathering data and performing regression analysis, Bozeman and Corley (2004) analyzed the collaboration among a group of scientists affiliated with universities in the U.S. and found a significant positive effect of funding on their collaboration. Using different independent variables, Adams *et al.* (2005) did another analysis and found that the researchers of top U.S. universities that have larger amounts of federal funding available tend to work in larger scientific groups that confirms the findings

of Bozeman and Corley (2004). Gulbrandsen and Smeby (2005) considered similar independent variables as Bozeman and Corley (2004) and Adams *et al.* (2005) to study the effect of industry funding on the performance of university professors in Norway. They used questionnaires to collect data from all tenured professors in Norway. In addition, they employed logistic regression rather than OLS linear regression. They observed that funded professors tend to collaborate more with other researchers from both universities and industries that confirmed the finding of the mentioned previous studies.

In another attempt on investigating the university-industry link, Lundberg *et al.* (2006) analyzed the effectiveness of co-authorship analysis in measuring university-industry collaboration and the impact of industrial funding on such collaboration. They focused on the industrial collaboration at Karolinska Institutet (KI), located in Stockholm (Sweden). Using indicators they compared the co-authorship data of KI with the industrial funding that was allocated to it. Their analysis includes 436 industrial companies that provided funding to the university. They found that two third of the companies co-authored at least one publication with the university. They also tried to confirm their findings from the companies' side and found that just 16% of the companies had co-authored publications. They concluded that their results are incomplete since they realized a conflict between the funding and co-authorship indicators.

Apart from all the above mentioned studies that mostly found a positive relation between funding and collaboration, Thune (2007) qualitatively analyzed the micro-dynamics of the collaboration among universities and industry using a social capital perspective. He concluded that social capital resources are important in forming collaborative projects. However, forming a successful collaboration is very difficult since it depends on a vast variety of other factors such as trust, familiarity, common language and understanding. In another study and employing logistic regression, Rosenzweig *et al.* (2008) analyzed the American papers published in the Academic Emergency Medicine, Annals of Emergency Medicine, Journal of Emergency Medicine, and American Journal of Emergency between 1994 and 2003. Although the collaboration as indicated by number of authors per paper increased during the examined period, they found no significant relation between collaboration and extramural funding.

Recently, Defazio *et al.* (2009) studied the impacts of funding on collaborative behavior and productivity of the researchers in the European Union (EU) funding program framework considering different funding periods in their analysis. In a 15-year period, they used a panel of 294 scientists in 39 EU research networks. They concluded that public funding may play an important role in forming more effective collaboration networks in EU region. The summary of the respective papers are depicted in Table 4.

Table 4. Summary of the research on evaluating the impact of funding on scientific collaboration

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
Beaver and Rosen	1979	Indicators	---	24 scientific fields	Positive relation
Heffner	1981	Statistics	· 395 articles · 28 journals · 1974-1975	4 scientific fields	Positive relation found in 2 (out of 4) disciplines
Bozeman and Corley	2004	Questionnaire Regression analysis	451 scientists and engineers in the U.S.	Scientists' collaboration choices	Significantly positive
Adams *et al.*	2005	Regression analysis	2.4 million papers 1981-1999	Top 110 U.S. universities	Positive effect of funding on team size
Gulbrandsen and Smeby	2005	Questionnaires Logistic regression analysis	1,967 records	Tenured university professors in Norway	A positive relation observed
Lundberg *et al.*	2006	Indicators	Industrial funding to a medical university 1993-2003	Co-authorship between university and industry	Incomplete results Some signs of positive effect on collaboration
Thune	2007	Qualitative analysis	Interviews with 29 researchers and R&D managers	Collaborative R&D projects in two academic fields	Important, but other factors are also involved
Rosenzweig *et al.*	2008	Logistic regression	5,728 articles published in 4 American journals 1994-2003	U.S. Emergency Medicine (EM)	No significant relation
Defazio *et al.*	2009	Regression analysis	Panel of 294 scientists 39 EU research networks	Researchers in the EU funding program framework	Funding may affect the formation of more effective collaboration networks

As it is shown in Table 4, researchers have started evaluating the impact of funding on the collaboration using simple indicators in the early 80s. Although their results mostly show a positive impact of funding, the approach that was used is not sufficiently rigorous to make any conclusion since they were mainly based on very simple indicators. In addition, their datasets were very limited. After a considerable time gap, researchers have started using more complex and integrated methods for analyzing the effect. One of the reasons could be

the availability of the digital data thanks to the progress in information technology. However, still some limitations in the methodologies and datasets can be seen. Lewison and Dawson (1998) did a statistical analysis on biomedical papers. One of the main limitations of their study is that they have considered the journal impact factor as their paper quality proxy instead of the number of citations which is a common practice in the literature. Using journal impact factor has several drawbacks (*e.g.* it is highly discipline dependent, editorial policies may affect the impact factor) and it is not accepted as a good paper quality measure (Moed, *et al.*, 1996; Seglen 1997). Bozeman and Corley (2004) used questionnaire (response rate of 45%) to gather their data and then did OLS regression on the collected data. The main concern about their study is the data since it is not representing all the university scientists and is very limited. This limitation can be also seen in Gulbrandsen and Smeby (2005).

Although Adams *et al.* (2005) studied a large collection of publications, they have not considered a time window to specifically analyze the effect of being funded in the current time window on the number of papers in the following year(s). In addition, the impact of funding has not been investigated at the individual level and it is limited to the university level. Among all the studies mentioned, Defazio *et al.* (2009) defined three different variables for funding reflecting pre-funding, during funding, and post-funding effects. However, they just focused on some well-known funded researchers.

Despite some studies that found a positive effect of funding, Lundberg *et al.* (2006) and Rosenzweig *et al.* (2008) could not find any significant relation between funding and collaboration. The reason may be that they have used a limited data source for their analysis. As an example, Rosenzweig *et al.* (2008) just considered the papers of 4 general peer-reviewed journals published in the United States. Thune (2007) used a qualitative approach for addressing the problem and using data from interviewing 29 researchers found a vague effect of funding.

As mentioned above, most of the studies employed statistical analysis. Table 5 shows important issues that have been considered as the crucial determinant factors in the studies that applied statistical analysis. As it can be seen, regional share and scientific field variables have attracted more attention of the researchers. Gulbrandsen and Smeby (2005) was the

most comprehensive study done from the number of independent variables and methodology points of view.

Table 5. Statistical analyses and variables, impact of funding on scientific collaboration

Article	Ctrl grp	Independent Variables*						Type of analysis	Sources of funding	Different funding periods	Network structure vars
		Fund	Gender	Prestige/ career /Age	Scientific fields	Regional share	Past productivity				
Heffner (1981)	✓	✓						Descriptive analysis	No	No	No
Bozeman and Corley (2004)		✓	✓	✓	✓			OLS Linear regression	No	No	No
Adams *et al.* (2005)		✓			✓	✓		Linear regression	No	No	No
Gulbrandsen and Smeby (2005)	✓	✓	✓	✓	✓	✓		Logistic regression	Yes	No	No
Rosenzweig *et al.* (2008)		✓			✓			Logistic regression	Yes	No	No
Defazio *et al.* (2009)		✓				✓	✓	Linear regression	No	Yes	No

* No of co-authors (or similar variables such authors/paper) has considered as the dependent variable

To be able to calculate the net impact of funding on collaboration several factors are required to be taken into consideration, *e.g.* control group of non-funded researchers, and different funding periods. If one neglects this kind of variables in the analysis then it would be hard to conclude that the resulting impact is directly due to the funding. According to Table 5, among all the reviewed studies just Gulbrandsen and Smeby (2005) considered a control group of researchers who applied for the funding but were not selected, but they have not included different funding periods. Heffner (1981) has also a control group but did not perform any regressions. Therefore, the net impact of funding on collaboration is still vague and it needs more analyses. For this purpose, availability of considerable digital datasets can help researchers to analyze the effect more comprehensively at the individual level. In the next section, the literature that has studied the impact of scientific collaboration on the output of the researchers is investigated.

2.4 Impact of Collaboration on Scientific Productivity

The modern science has become more complex and interdisciplinary in its nature. This encourages researchers to be more collaborative and get engaged in larger collaboration networks (Lee & Bozeman, 2005). Analyzing the impact of collaboration on scientific productivity is not a recent issue. Several pioneer collaboration studies mentioned that collaborative activity increases the productivity of the researchers (Lotka, 1926; De Solla Price & Beaver, 1966; Zuckerman, 1967). Lotka (1926) was the first one who studied the

productivity of the researchers and the impact of the team size. He analyzed the scientific productivity distribution in physics (since the beginning of "*Auerbach´s Geschichtstafeln der Physik*" journal publication until 1900) and chemistry (1907-1916). His data was very limited especially for the chemistry journal articles where he only considered the authors whose surnames began with the letters A and B. He found that there exists an inverse relation between the number of articles and the number of the co-authors that produced them. After a considerable time, this topic has been examined again in 1960s. De Solla Price and Beaver (1966) studied the publications and collaboration of 592 researchers and found a good correlation between their collaboration and scientific output. They realized that the scientist with the highest number of publications was also the most collaborative one. In 1967, Zuckerman interviewed 41 Nobel laureates and found a strong correlation between their productivity and collaboration. In general, they published and collaborated more than normal researchers. Among these pioneering articles, the work of De Solla Price and Beaver (1966) seems more interesting since he used co-authorship as a proxy for collaboration and highlighted the ongoing growth of the scientific collaboration.

It is very likely to assume that scientists benefit from the collaboration to increase the quantity and quality of their scientific output through being involved in larger research teams and having better access to resources (Katz & Martin, 1997; Melin, 2000; Beaver, 2001; Heinze & Kuhlmann, 2008), to expertise (Katz & Martin, 1997; Thorsteinsdottir, 2000) and to funding (Beaver, 2001; Heinze & Kuhlmann, 2008). Through scientific collaboration researchers interact with each other and can criticize team members' work (or duties). This internal referring may thus result in a higher quality publication (Salter & Martin, 2001; Lee & Bozeman, 2005; Adams, *et al.*, 2005).

Lee and Bozeman (2005) did a statistical survey study of 443 academic researchers in the United States. They used two-stage least square method to analyze their collected data. They defined several independent variables such as age, rank, gender, job satisfaction, *etc.* and found that the number of peer-reviewed journal articles of the researchers is significantly related to the number of their collaborators. However, they emphasized that the net impact of collaboration is still less clear. Moreover, they just focused on the individual level of collaboration and neglected the benefits of collaboration that might be gained from scientific groups, institutions or scientific fields. In another study, Adams *et al.* (2005) studied the

scientific collaboration and team size in 110 top universities of the United States over the period of 1981-1999. They considered number of authors per paper as a measure of scientific team size. Their results show that the average team size has increased by 50% over the examined period. Moreover, they concluded that quantity and quality of the scientific output increase with team size and scientific collaboration.

Pao (1982) focused on the field of computational musicology. He realized that more collaborative musicologists were more productive. However, just 15% of the publications in musicology were published in collaboration with other authors. Pravdic and Oluic-Vukovic (1986) analyzed the collaborative behavior and the publication rate of the researchers involved in chemistry. They identified a close relation between the number of the articles of the researchers and their collaboration pattern. They proposed that if a scientist is highly productive (low-productive) then collaborating with him/her might increase (decrease) the number of publications.

Collaboration enables researchers to share special, expensive and unique equipments (Meadows, 1974; Thorsteinsdottir, 2000) that may help them to increse their productivity. Thorsteinsdottir (2000) studied and compared external research collaboration in two small regions, Iceland and Newfoundland (Canada). He employed bibliometric descriptive analysis and interview data to asses the collaboration quantitatively and quallitatively. The bibliometric data was collected from the Science Citation Index (SCI) database for the period of 1990-1994. His results show that apart from having a better access to funding sources, researchers in the mentioned regions do collaborate to share the research material and equipment.

Collaboraiton makes it possible to mentor univeristy students (Beaver & Rosen, 1978, 1979) that may lead to the enhanced productivity of individual researchers (Melin, 2000). Melin (2000) focused on the reasons for collaboration at the individual level and analysed the interactions within research teams. His data consisted of 195 records that were collected through sending questionnaires to all first-listed authors who were in the 1994 CD-ROM version of Science Citation Index (SCI) and affiliated with the Umea University in Sweden. According to his findings, 14% of the respondants believe that supervisor-student realtion is the major reason for the collaboration to publish an article.

Most of the studies that analyzed the impact of collaboration on scientific productivity are limited in scope. Bordons *et al.* (1996) focused on three biomedical areas (neurosciences, gastroenterology, and cardiovascular systems) and employed bibliometric analysis to analyze the impact of collaborations on scientific productivity. They found that international and intramural collaborations have a positive influence on the productivity of individual authors. This might be due to the reason that through collaborations scientists may have the opportunity to work on different projects simultaneously. They observed that researchers who work in applied science tend to collaborate locally whereas researchers working in basic science prefer internal collaboration aiming to publish in higher quality journals.

Martin-Sempere *et al.* (2002) studied the impact of intramural and extramural collaboration on productivity. They found that researchers who belong to no scientific group show lower productivity and they tend less to collaborate internationally. These results are expected since researchers with no group generally have lower access to funding resources. Through collaboration they can be involved in more projects. In another study, Mairesse and Turner (2005) studied the impact of collaboration among researchers who worked at the French Centre National de la Recherche Scientifique and found a significant positive impact on scientific productivity. Employing bibliometric indicators on the data of astronomical research in the Netherlands, van Raan (1998) studied the impact of international collaboration on the quality of the research output and found a positive relation.

In a recent macro-level study, Tang and Shapira (2012) used bibliometric analysis and statistical testing to analyze the effects of international collaboration on China's nanotechnology research impact. They just studied the impact on the quality of the papers and supposed that collaboration effect on scientific productivity is positive. The main goal of their study was finding out if with the raise of Chinese publications through domestic and international collaboration, their quality has also improved. They found that Chinese researchers who bridge scientific worlds by publishing scientific papers with both domestic and international scientists have a positive impact on the quality of Chinese articles.

Nevertheless, there still exist arguments about the relation between scientific collaboration and productivity indicating that the evidence in the literature is contradictory (Lee & Bozeman, 2005). For example, there are some suggestions that collaboration has in

fact a negative impact on scientific productivity. As mentioned in the previous section, working in research teams may cause transaction costs (Landry & Amara, 1998). Moreover, working with others needs more time and energy since it is required to have more communication, waiting for other comments or even waiting for a member to do his part of job. University-industry collaboration can also have some side effects. Nelson (2004) argues that collaborating with industry might delay or prevent scientific publication since the industrial partners may prefer not to expose the results.

Banal-Estanol *et al.* (2008) focused on the researchers from the engineering departments of 40 major universities in the UK and studied the impact of university-industry collaboration on their scientific productivity and on the research itself during 1985-2007. Performing regression analysis, they concluded that researchers who are in collaboration with the industrial companies publish significantly more articles in general. However, the industry collaboration has a negative impact on the number of basic research publications and it resulted to more applied articles. In other words, basic research suffers from industrial collaboration and industrial links can change the direction of the research. This issue has been also investigated in two other studies that are based on questionnaire data. Blumenthal *et al.* (1986) and Gulbrandsen and Smeby (2005) showed that industry support may influence the choice of the topics that academic researchers select to work on.

Therefore, not all the collaborations will result in higher productivity. It is very likely in the scientific communities that an active collaborator has un-finished project(s) due to the low performance of one (some) of the team members (Lee & Bozeman, 2005). Senior researchers tend to collaborate less (Bozeman & Corley, 2004), which may be explained by a lower motivation to increase their own productivity at a later stage of their career. Or, experienced researchers are worried to lose their productivity by involving other un-experienced researchers (Lee & Bozeman, 2005).

Godin and Gingras (2000) analyzed the impact of collaborative research on the scientific publication of Canadian researchers indexed in the Science Citation Index (SCI) during 1980-1997. One of their main research questions was to investigate if collaboration has a negative impact on scientific productivity due to limitations that collaboration may impose on the time and resources of the researchers. They divided the Canadian publications

into two general categories; those that were published in collaboration with local partners and those that were published collaborating with international ones. They argued that the collaboration does not empirically harm the scientific productivity of the examined Canadian researchers but they recommended monitoring the situation continuously. In addition, they observed a higher growth rate for international collaboration in comparison with local collaboration among the examined Canadian researchers. In the section, the research gaps are discussed.

2.5 Research Gaps

Measuring and understanding the effect of funding is very critical. Financial investment may not necessarily be effective and may not result in higher scientific productivity. In order to have an efficient funding allocation, we need to understand the determinants of productive investment. If we know the marginal impact of research funding across disciplines, universities, researchers *etc.* it would be easier to plan a more efficient system. Based on the literature reviewed, the relation among funding, collaboration and scientific production is still vague. The relation between collaboration and scientific productivity is not clear and the results of different papers are contradictory. Especially, our knowledge about the effects of funding on collaboration is very limited. Most of the studies considered a very limited scope such as the collaboration among the university professors or the cooperation between universities and industry. Moreover, no testing group has been considered in most of the studies. The other important issue is that we rarely found comprehensive analysis in the papers and whenever we see for example statistical analysis they are in the form of simplified linear regressions including limited number of variables and a narrow data set. It is thus suggested to consider other forms of regressions such as non-linear equations or to add cross-relations between the independent variables to analyze the combined effects of the independent variables while benefiting from the availability of the data sets of considerable sizes.

Network structure variables are important factors in evaluating the effect of collaboration or analyzing the collaboration patterns. Considering such variables help us to study the impact of scientific collaboration more accurately. Although this issue was recently considered by Beaudry and Allaoui (2012), they have not focused on the effect of network

structure variables and did not come with a comprehensive picture of the role of the network architecture. Moreover, in order to better analyze the impact of funding it is worth to consider the network variables in combination with other determinant factors (like past productivity of the researchers) to study the impact more precisely.

To my knowledge, the role of funding and collaboration in scientific productivity has not been examined for the prominent individuals like star scientists and gatekeepers. Detecting the researchers who are playing an important role in scientific production and collaboration, and analyzing their collaboration trends and funding patterns seems necessary in order to better understand their role in stimulating scientific activities and enhancing research productivity in their communities. Such analysis will enable us to discover the characteristics of important players in scientific collaboration networks and will help policy makers to align their strategies in a way that improves the overall efficiency of the funding programs and collaboration networks. Nature of the science is becoming more complex and inter-disciplinary. Expertise from more and more disciplines is becoming necessary in order to produce knowledge. This should be reflected in funding policies of the governments and granting agencies, but no research so far has developed a clear link between the funding, multidisciplinary collaboration and knowledge production.

Most of the studies have used bibliometrics or statistical methods for performing the analysis. Although bibliometrics is a simple and easy to use method, it is not an integrated approach since it considers too many assumptions that make the model very simplified (Ruegg, 2007). This could be also true for limited scope statistical analysis and econometrics since the model is very limited and simplified in comparison with the real problem (Salter & Martin, 2001). Therefore, it is suggested to employ a variety of techniques such as data and text mining, social network analysis, bibliometrics, statistical analysis, and visualizations to complement and validate the findings.

It can be said that the most comprehensive work that has been done in evaluating the impact of funding on scientific productivity of Canadian researchers is Godin (2002). However, it just considered the impact of funding on the scientific productivity and neglected the impact of collaboration or network structure. Bibliometric indicators have been used in this paper and no integrated statistical validation is given. And, the analysis and data

back to the period of 1990-1999 that indicates the importance of an in-depth comprehensive evaluation at the individual level of the researchers while covering more recent data and utilizing more techniques. Considering the above, the main purpose of this thesis is to employ various techniques and large data sets to do a comprehensive study and an integrated evaluation at the individual level of the researchers. This can definitely help Canadian policy makers to adjust their strategies in a way that will lead the country to a better scientific position worldwide.

3 RESEARCH QUESTIONS AND OBJECTIVES

In this research, the inter-relations among three main target variables (*i.e.* funding, scientific output, and collaboration) are studied. The focus is on the inter-effects of NSERC funding, funded researchers' collaboration and output in natural sciences and engineering in Canada within the period of 1996 to 2010.

3.1 Research Questions

Several questions are addressed in this thesis:

- Scientific output:
    1. What is the impact of funding on quantity and quality of the scientific output?
    2. Does it have the same impact in different scientific disciplines?
    3. Does funding increase the performance of the scientists and make them more productive?
    4. How productive are Quebec researchers in comparison with the other provinces of Canada?
    5. What are the most (least) productive Canadian provinces?
    6. How different Canadian universities are acting against the funding they receive?
    7. Are researchers with higher level of funding more productive?
    8. What is the best measure for quantifying quality of the research?
    9. Is publications' quality affected by the level of funding?
    10. What are the characteristics of the researchers who usually produce high quality papers?
    11. How were the trends of quantity and quality of the Canadian funded researchers in the past fifteen years?
    12. What is the impact of collaboration on quantity and quality of the scientific output?
    13. What is the best model for predicting the productivity of the funded researchers?
- Scientific collaboration:
    1. Does funding have an impact on the collaboration pattern among scientists?
    2. Can funding affect the position of the researchers within their collaboration network?

3. Is there any relation between the profile of the researchers and their collaboration patterns?
4. How different Canadian provinces' researchers collaborate?
5. What is the trend of collaboration in top Canadian universities?
6. Does NSERC funded researchers' collaboration network resemble the "*small world*" property? If yes, what is the impact of the property on scientific activities of the funded researchers?
7. What is the impact of scientific profile of researchers and their level of funding on their positions within the collaboration network?

- Funding:
  1. How NSERC allocates funding to different Canadian provinces?
  2. How researchers can earn more funding?
  3. How NSERC allocates funding to the top Canadian universities' researchers?
  4. Does being more collaborative result in higher amount of funding?
  5. Is there any relation between the past productivity and the amount of funding in the following year?
  6. How profiles of the researchers affect their level of funding?
  7. Is there any relation between position of the researcher in the collaboration network and the amount of funding that he/she receives?

According to the literature reviewed and the above mentioned questions, several objectives are defined. In this section, the goals of the research are discussed in detail.

3.2 Research Objectives

The general objective and specific objectives of the research are presented in this section.

3.2.1 The General Objective

The general objective of this research is to investigate the inter-relations among funding, scientific activity, and collaboration. For this purpose, the impact of the funded research on the scientific development and researchers' performance in terms of the quantity and the quality of their output and its effect on collaboration rate and patterns of NSERC

funded scientists. In addition, the impact of scientific collaboration on the productivity of the researchers and their level of funding is investigated. Moreover, the impact of the most determinant influencing factors of researchers' funding is evaluated.

3.2.2 The Specific Objectives

The specific objectives of this research are as follows:

- Apply a unique approach:
    1. To extract the data (*i.e.* publications, funding, annual citations) from the internet-based sources and create the target data automatically
    2. To calculate the bibliometric indicators
    3. To calculate the network structure variables
    4. To match the same records in different databases
    5. To clean the data and detect outliers
    6. To integrate all the collected data in a single database
- Validate all the research assumptions:
    1. Through holding interviews with researchers of different categories
    2. Sending questionnaire to the respondents that are selected based on stratified sampling method
- Factors affecting scientific output:
    1. Examine effect of funding on article quantity and quality (at the level of individuals, provinces, research universities, major scientific areas, career status, demographic variables, and impact of various funding programs)
    2. Evaluate the impact of scientific collaboration on quality and quantity of the scientific output (at the level of individuals, provinces, research universities, major scientific areas, career status, demographic variables, and different collaboration network's positions)
    3. Determining the critical factors of scientific productivity, *i.e.* what are the determinant factors of productive researchers?
- Factors affecting scientific collaboration:

1. Check if the NSERC funded researchers' collaboration network resembles the small world property and assess the impact of the property on the collaboration patterns and scientific productivity of the researchers
2. Investigate the impact of the influencing factors (*e.g.* funding, researchers' profile, past productivity, demographic variables) on the group structure and scientific collaboration (in terms of size of the research groups, multi co-authorship patterns, researchers' position in the collaboration network) at the individual level of the researchers

- Factors affecting funding:
  1. Determine the impact of collaboration patterns and productivity of researchers on funding
  2. Determine the most determinant factors for the researchers to get higher amount of funding
- Machine learning classification and prediction framework:
  1. Using the results from other methodologies, suggest a reliable highly accurate machine learning model for:
     - Classifying the funded researchers based on their scientific profile and funding level
     - Predicating the productivity of the researchers
     - Proposing an approximate funding that a researcher is deserved to get in a given year based on his profile and past productivity
- Policy Implications
  1. Make recommendations for policy makers, in terms of the efficiency of various funding programs on performance of the researcher, the distribution of money among funding categories, the most efficient researchers, institutes, provinces, *etc.*

4 DATA AND METHODOLOGY

4.1 Data Extraction

Initially, this research required the NSERC funding data to be collected for the period of 1996 to 2010. As the next stage, all the publications that have been produced by the NSERC funded researchers within the mentioned time interval were extracted. Finally, all the collected data were integrated into a single database and another numerical database was generated to be used for the data mining analysis. The data gathering procedure is described in detail in this chapter and the overview of the methodologies is presented.

4.1.1 NSERC Funding Database

NSERC was selected as the source of funding in this research since it is the main federal funding organization in Canada. Almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC (Godin, 2003). As the first step, a JAVA program was coded to collect the funding related information (*e.g.* grantee, funding program, title of the award, grantee's affiliation, year of award, amount, *etc.*) from the NSERC public database within the period of 1996 to 2010 and integrate the result into a single database, named *fundingDB*.

4.1.2 Publications Database

Creation of the publications database was involved with two main challenges, *i.e.* to select the most suitable source of scientific publications for the analysis, and to detect the papers of the NSERC funded researchers. This section discusses the data gathering procedures and the respective assumptions in detail.

4.1.2.1 NSERC Funded Researchers' Articles

To collect the publications data the common procedure used in the literature is to list the funded researchers and then to collect all the articles that were published by the funded researchers. This will surely result in an overestimation of the number of publications since a researcher can have various sources of funding at the same time. In other words, suppose we have a NSERC funded researcher named *A*. If we collect all the articles published by *A* there will be surely some articles that were supported by other sources of funding (not necessarily NSERC). Hence, to overcome the mentioned limitation I made an assumption that all the

funded researchers should acknowledge the source of funding in their articles (which is the case based on NSERC's regulations). To validate the assumption, 4,000 of the NSERC researchers in the *fundingDB* were randomly selected and a questionnaire was sent to them to see if the assumption is valid. 401 researchers responded to the questionnaire from which 89.3% confirmed that the source of funding is acknowledged in the publications. Hence, based on the defined approach the publication data source (is explained in 4.1.2.2) should provide a full text search to check if the support of NSERC is acknowledged in the paper.

4.1.2.2 Publication Data Source

As the first stage, various digital scientific publications sources were compared based on various criteria, *e.g.* full text search ability, number of publications covered, number of publishers, authors' affiliation information, abstract of the articles, accurate and comprehensive meta data, *etc*. In addition, some scientific search engines (*e.g.* Google Scholar, Microsoft Academic Search) were also considered and compared. The reason for considering the scientific search engines was the high coverage of the publications. However, they suffer from some problems, *e.g.* dead links, and inconsistent accuracy (Falagas, *et al.*, 2008).

Comparison of various data sources as well as the data requirements of the research was led to a new data extraction methodology that benefits from the advantages of the sources. Specifically, Elsevier's Scopus was selected as the main source of the publications data since it provides comprehensive and highly accurate information especially after 1996. In addition, Google Scholar search engine was used since it provides the full text search over publications. Combining Google Scholar with Scopus has some advantages, *i.e.* using Scopus authors' id, getting access to the history of an author's affiliations in Scopus, indexed keywords of the articles in Scopus, consistency and high quality of the data in Scopus, benefitting from wide-scale full text search of Google Scholar.

The period of 1996 to 2010 was selected as the time interval of the research since Scopus data quality was higher than 1996 and articles needs at least three years to be cited hence I stopped at 2010. According to the defined methodology, first publications were searched over Google Scholar within the period of 1996 to 2010 and the ones that

acknowledged NSERC funding were listed in separate text files for each year. As the second step, text files were fed into a JAVA program to automatically collect all the required information of the listed articles from Scopus. Scopus provides total citation counts of the publications. However, annual citation data of the articles were needed to assess the quality of the papers more accurately. SCImago was selected for collecting the impact factor information of the journals in which the articles were published in as well as annual citation counts of the papers. A JAVA program automatically collected the required information. SCImago is powered by Scopus that makes it more compatible and consistent with the publications database. The results were integrated and stored in a MySQL database named *pubDB*.

4.2 Data Cleaning and Integration

4.2.1 Data Cleaning

After collecting the data, *fundingDB* and *pubDB* were cleaned extensively. Irrelevant or missing data, empty data fields, non-English characters, and splitting the affiliations of the authors were some the most frequent problems. For this purpose, a JAVA program was coded and used to clean the mentioned databases separately. After the automatic cleaning procedure, the data was checked randomly to detect the problematic issues. This recursive procedure of automatic cleaning and random check was performed several times.

4.2.2 Data Integration

After cleaning the databases, *fundingDB* and *pubDB* should be integrated. In other words, the funded researchers in *fundingDB* were needed to be identified in *pubDB* and get a similar ID as the one in *pubDB*. This was a very challenging issue since it involved with disambiguation of the entities in *fundingDB* and *pubDB*. For this purpose, a JAVA program was coded that compared each of the funded researchers with the records in *pubDB* based on various criteria (e.g. first name and last name of the author, affiliation, research area, *etc.*) and calculated a similarity probability. If the similarity probability was higher than 90% it automatically assigned the ID of the author in *pubDB* as ID of the funded researcher in *fundingDB*. If the probability was lower than 90% and higher than 50% the program asked for the user input to confirm if the records were the same. Otherwise, it disregarded the

record in *pubDB* and took another record. The result of this time consuming procedure was an integrated database named *nsercDB*. A secondary database was generated from *nsercDB* by calculating various bibliometric and network structure indicators. The resulted numerical features along with other required information (*e.g.* demographic variables) were integrated into another database named *miningDB* that was used for the data mining analysis.

4.3 Methodologies and Tools

In general, this research employed a triangulation of the following methodologies and methods:

- Visualization techniques
- Bibliometrics
- Social Network Analysis (SNA)
- Statistical analysis
- Data and text mining
- Survey data analysis

Predictive and descriptive visualizations were made by means of RapidMiner data mining package, STATA statistical analysis software, and Microsoft EXCEL 2010. Visualizations helped to better understand the data features and behavior while providing insights over the data. In addition, the resulted graphs and diagrams were used to recognize primary and easy-to-detect patterns. By means of bibliometric indicators data was analyzed rigorously. Major aspects and behaviors of the data were examined over the defined fifteen year time span (1996-2010) by means of a large variety of bibliometric indicators. Several statistical models were defined and tested over a number of variables to evaluate the inter-relations among funding, collaboration, and scientific output. To validate the assumptions of the research and to check the results, a questionnaire was designed and sent to the target respondents and the responses were statistically analyzed. As the final stage of the research, a machine learning framework was designed and suggested for evaluating the performance of the researchers and proposing their deserving amount of funding in a given year. Since this is a manuscript-based thesis, more detailed information about the mentioned

methodologies and methods is given in each of the papers in Chapter 5 separately, whenever applied. In the following chapter, the results are presented and discussed.

5 RESULTS

In this section, the results of the research are presented and discussed in four sections. First, the results of the bibliometric analysis are presented. The next section is dedicated to the statistical analysis results while the third section presents the machine learning framework. The chapter concludes with the survey data analysis. In total, nine papers are discussed in this section. Two other publications that were produced at the initial stages of this research are listed in Appendix A.

5.1 Bibliometrics

Three papers were produced based on bibliometrics method that are presented in separately in this section. The first paper, titled "*Bibliometric Analysis of the Impact of Funding on Scientific Activities of Researchers*", evaluates scientific productivity and collaboration patterns of the NSERC funded researchers residing in different Canadian provinces and is presented in section 5.1.1. Section 5.1.2 discusses the results of the second paper, titled "*Investigating Scientific Activities in Various Disciplines and the Impact of different Funding Programs*". This paper compares the effect of various NSERC funding programs on scientific activities and funding of researchers. In addition, it evaluates funding, scientific productivity, and collaboration patterns and their inter-relations in different scientific disciplines. Section 5.1.3 belongs to the third bibliometric paper, titled "*Analyzing Scientific Activities of the top Ten Canadian Universities*". This paper focuses on the scientific activities of researchers who are affiliated with the top ten Canadian universities in 2013 and evaluates their funding and scientific performance patterns within the period of 1996 to 2010.

5.1.1 Bibliometric Analysis of the Impact of Funding on Scientific Activities of Researchers

Funding has been acknowledged in many articles to be the main determinant of scientific development and it is viewed as an important factor having a significant effect on the scientific output. Every year, a considerable amount of money is being invested on research, mainly in the form of funding allocated to universities and research institutes, in order to improve the scientific potential of the country. Hence, to better distribute the

available funds and to set the most proper R&D investment strategies for the future, evaluation of the productivity of the researchers in respect to the amount of funding that they have received and the impact of such funding is crucial. In this paper, using the data on 15 years of journal publications of the funded researchers and by means of bibliometric analysis [6], the scientific output of the researchers and their collaboration patterns is investigated. Our focus is on the Canadian researchers who are active in the field of natural sciences and engineering and reside in different Canadian provinces. According to the results, funding has a different impact in low and high funding Canadian provinces. In high funding provinces funding mainly affects the quality of the works while in low funding group of provinces it has some impact on the rate of publications but not on the quality of the papers, measured by average number of citations. However, no relation was found between funding and the average impact factor of the journals in which researchers publish their articles. It was observed that funding influences the scientific team size of the researchers in all the Canadian provinces.

5.1.1.1 Introduction

Scientific activities and size and quality of the R&D sector play a key role in determining the world-wide position of a country. Many articles have acknowledged funding as the main determinant of research productivity (*e.g.* Martin, 2003; Boyack & Borner, 2003; McAllister & Narin, 1983) and the level of funding has been indicated as the most critical factor for improving the research productivity. Although the approach towards the allocation of research funding varies across the countries, and different strategies and procedures are being followed worldwide for this purpose, governments are annually investing considerable amounts of money in R&D in a hope to stimulate a higher scientific performance of the funded researchers.

It is easy to judge productivity and impact of research of the Nobel laureates or star scientists (extremely productive scientists). However, for the rest of scientists one should have quantitative indicators to analyze and compare the scientific productivity of the researchers (Hirsch, 2005). Publications are usually considered as the main output of the scientific activities (*e.g.* Drummond, 1997; Naoki, 2008). They are also viewed as the

[6] Employing a set of methods to analyze the academic literature quantitatively (De Bellis, 2009).

principal measure of academic recognition in most of the western countries (Horton, 1998). It is claimed that only a limited number of journal papers is currently publishing the main results of the scientific research (Shibata *et al.*, 2008). In addition, a small number of scientists are publishing most of the scientific papers and the weights of publications are not distributed evenly (Bookstein, 1980). This is known as the Lotka's law in the literature, introduced by Lotka (1926).

Governments have funded research for more than sixty years (Godin & Doré, 2004) and have employed various tools and techniques, both quantitative and qualitative, to measure their scientific performance (Godin, 2002). Having such a history, the impact of funding on the scientific output has been investigated in the literature from various perspectives. A few studies assessed the impact of funding on the productivity of the medical schools or programs (*e.g.* McAllister & Narin, 1983; Lewison & Dawson, 1998; Albrecht, 2009). A number of studies focused on the effect of contractual funding on the quantity and quality of the scientific publications (*e.g.* Arora & Gambardella, 1998; Carayol & Matt, 2006). Using statistical analysis, various studies investigated the impact of federal funding (*e.g.* Payne & Siow, 2003; Huffman & Evenson, 2005), industry funding (*e.g.* Gulbrandsen & Smeby, 2005), or private funding (*e.g.* Beaudry & Allaoui, 2012) on scientific productivity and research performance. In addition, a few studies focused on the scientific productivity at the countries level and assessed the impact of national investments (*e.g.* Leydesdorff & Wagner, 2009; Crespi & Geuna, 2008). For a complete survey on the topic see the comprehensive literature review of Ebadi and Schiffauerova (2013), listed in Appendix I.

Evaluating the impact of funding has also attracted the attention of the Canadian researchers. In the studies evaluating this effect in Canada, scientific articles have been considered as the main output of researchers and universities (Godin, 2003) and bibliometrics has been mostly used for scientific evaluation purposes. Using data for the period of 1984-1993, Gingras (1996) in a report to the Program Evaluation Committee of Natural Science and Engineering Research Council (NSERC) discussed the feasibility of bibliometric evaluation of the funded research. Godin (2003) in a bibliometric evaluation studied the impact of NSERC funding on the productivity and papers' quality of the supported researchers for the period of 1990-1999. He used two simple bibliometric

indicators over Science Citation Index (SCI) database and analyzed the number of papers written by funded researchers over a 10-year time period to find NSERC's contribution to the scientific development of Canada. He found that researchers with higher amount of funding available are more productive. In addition, when the level of funding for a given researcher is above the median (high) his/her productivity is more strongly correlated with the amount of funding. However, the level of funding does not affect the researchers' journals quality.

In a series of case studies, Campbell and his colleagues performed bibliometric evaluations on the impact of funding on scientific performance (Campbell, *et al.*, 2010; Campbell & Bertrand, 2009; Campbell, *et al.*, 2009). In two recent studies, Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) used multiple regression analysis to study the impact of public and private funding on the scientific production of the Canadian academics working in biotechnology and nanotechnology fields respectively. They found a positive impact of public funding and no impact of private funding on number of publications.

Apart from Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012), studies that evaluated the impact of funding in Canada used limited and simple indicators. In addition, the size of the dataset and the scope of the research are limited and their results are thus not necessarily valid outside the defined scope. Moreover and as discussed, most of the studies assessed the impact of funding on the scientific output in terms of the rate of publications but not the quality of the papers. However, funding can also affect other aspects of the scientific activities. As an example, it may influence the scientific collaboration patterns among the researchers that may result in higher/lower scientific productivity, which is one of the interests of our work.

The objective of this work is to evaluate the impact of funding in natural sciences and engineering in Canada on the scientific production of the funded researchers and on their collaboration patterns. This paper is more comprehensive than most of the existing research as it involves all the engineering and natural sciences researchers within the whole country. Our work extends the literature in four ways. First, we will use a larger and more recent data set spanning from 1996 to 2010 that will be defined in detail in the section 5.1.1.3. Secondly, apart from analyzing productivity and quality of the work of the researchers, it also assesses

the impact of funding on scientific collaboration patterns. Thirdly, it evaluates the impact of funding on scientific activities of the researchers while focusing on different impact in different Canadian provinces. And finally, we use a unique procedure for finding the articles that have acknowledged the source of funding in the body of the paper. This is a crucial step in assessing the impact of funding that has been neglected in the previous studies. The common procedure in the literature is counting all the articles that have been published by a funded researcher which creates a great bias (overestimation). However, we will only count those that have really acknowledged the source of funding. The procedure will be discussed in detail in section 5.1.1.3. The rest of the paper is organized as follows: Section 5.1.1.3 describes methodology and data that will be used in this study. The empirical results and interpretations are provided in section 5.1.1.4. Section 5.1.1.5 presents the findings of this research and the limitations of this study and some directions for the future work are discussed in 5.1.1.6.

5.1.1.2 Data and Methodology

NSERC was selected as the funding organization of interest since it is the main federal funding organization in Canada. Almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC (Godin, 2003). In our study, we focus on the period of 1996 to 2010. The reason for choosing 1996 as the beginning year of the analysis was better coverage of Scopus after 1996. The lower quality of the data before 1996 might affect the results. Hence, we first collected the funding data from NSERC for the period of 1996 to 2010 that contained information like name of the researcher, his/her affiliation, year, and amount of the award. The extracted data was then refined by employing several automatic cleaning modules coded in Java.

In addition, team grants were associated to the principal investigator in the original database where we divided the amount equally among the researchers of the team. In order to confirm such assumption we held several interviews with researchers selected from our database, where most of them supported the validity of such approximation. The final refined funding dataset contains 75,967 distinct Canadian researchers who received funding from NSERC during the aforementioned period.

As the next step, we searched over Scopus to gather the articles of the NSERC funded researchers for the mentioned period. For this purpose, we searched for all the articles that had acknowledged NSERC funding support within the body of the article. This was a very crucial step in fetching more accurate data that will highly influence the findings of the research. The common procedure in similar studies is finding all the articles of the funded researchers that may result in an over estimation, especially if there is a great variety of the funding sources. Our procedure is based on the assumption that comes from the NSERC guidelines which stipulate that the funding received through NSERC has to be acknowledged in each supported article of the funded researchers. Hence, by our procedure we only take into consideration the articles that were produced as the result of NSERC funding, not all the articles of the researcher. We assume that this certainly leads to a more accurate data and analysis. All the related information such as article co-authors, co-author affiliations, article title, abstract *etc.* was then extracted. The articles dataset contained in total 130,510 articles and 177,449 authors that acknowledged NSERC support in the respective article.

For evaluating the quality of the papers, SCImago was selected for collecting the impact factor information of the journals in which the articles were published in and the result was integrated into another dataset. SCImago was chosen for three main reasons. First, it provides the journal impact factors for each of the single years of our examined time interval. This enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published and not its impact in the current year. The impact factors evolve, and there may be a significant difference between the impact factor of the journal in the current year or the same measure in 15 years ago. The respectability of the journal and consequently the quality of the published articles are best judged by the impact factor in the year the article was published.

The second reason is related to the coverage and quality of SCImago as an open access resource. According to Falagas *et al.* (2008), SCImago covers considerably more journals in comparison with Web of Science which serves as the basis for calculating journals' more commonly known Impact Factor (IF) published by Thomson Reuters. In addition, SCImago contains a wider variety of countries and languages. Moreover, contrary to Thomson

Reuters's IF, SCImago's SJR indicator uses different weights for citations depending on the quality of the citing journal. Due to the mentioned advantages, SJR indicator is now considered as a serious alternative to Thomson Reuters's IF. Finally, SCImago is powered by Scopus that makes it more compatible with our articles database.

Having all the required data collected, we search for relationships between the amounts of funding that NSERC has allocated to the researchers and their scientific productivity in terms of the number of publications and quality of the papers. In addition, the impact of funding on the collaboration patterns of the researchers is analyzed. Bibliometric analysis is used for this purpose to assess the scientific productivity and collaboration patterns of the funded researchers.

5.1.1.3 Results

The scope of the analysis is Canada-wide and the impact of funding is evaluated for different Canadian provinces. Using bibliometric indicators, the productivity and collaboration of the Canadian researchers are compared. In the literature, three-year (*e.g.* Payne & Siow, 2003) or five-year (*e.g.* Jacob & Lefgren, 2007) time window has been considered for funding to show effects on scientific activities. In this research, we consider three-year time window for publications of the funded researchers. One-year time window for the scientific output of the researchers was also included in the analysis in order to assess the impact in the year of funding and compare the results with the three-year time window. As an example of the three-year time window, if the year of funding for a researcher is 1996, we gathered all his articles that acknowledged NSERC funding for the period of 1996 to 1998.

We considered NSERC funded researchers from all the ten Canadian provinces. We excluded Canadian territories (*i.e.* Yukon, Nunavut, and Northwest Territories) from our analyses since the calculated indicators were too small for the mentioned territories in comparison with the ones for provinces. In addition, we also excluded student funding programs. In the rest of this section, we discuss the results in four separate sections which are: funding, funding and rate of publications, funding and publication quality, and funding and team size.

5.1.1.3.1 Funding

First we analyzed the amounts of funding in each province. As it can be seen in Figure 5, Canadian provinces can be divided into two groups based on their total share from NSERC funding. The first group contains Ontario, Quebec, British Columbia, and Alberta that have received considerably higher share of NSERC funding from the provinces of the second group. Saskatchewan, Nova Scotia, Manitoba, New Brunswick, Newfoundland & Labrador, and Prince Edward provinces belong to the second group that have received comparable but much lower total share of funding than the provinces in the first group. We will use the terms "*high funding provinces*" and "*low funding provinces*" in the rest of the section for pointing to the aforementioned provinces.

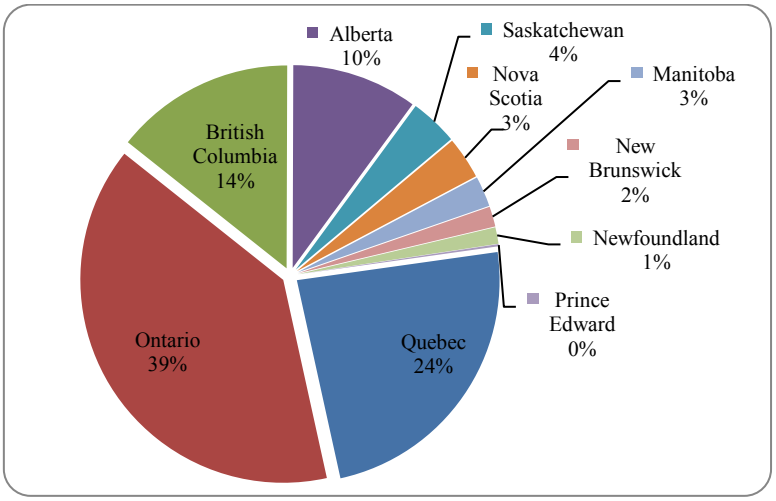


Figure 5. Total funding share of Canadian provinces, 1996-2010

Although there are considerable differences in the total amount of NSERC funding allocated to the Canadian provinces, the average amounts of funding dedicated to the researchers are quite comparable. According to Figure 6-a, the average total amount of funding per researchers in the examined provinces was in the range of 8-13 percent. More interestingly, this share is the same for all the members of the high funding provinces, having the level of 11 percent. Moreover, although Ontario had the highest level of total funding with a considerable difference, Saskatchewan is the highest if we consider the average share.

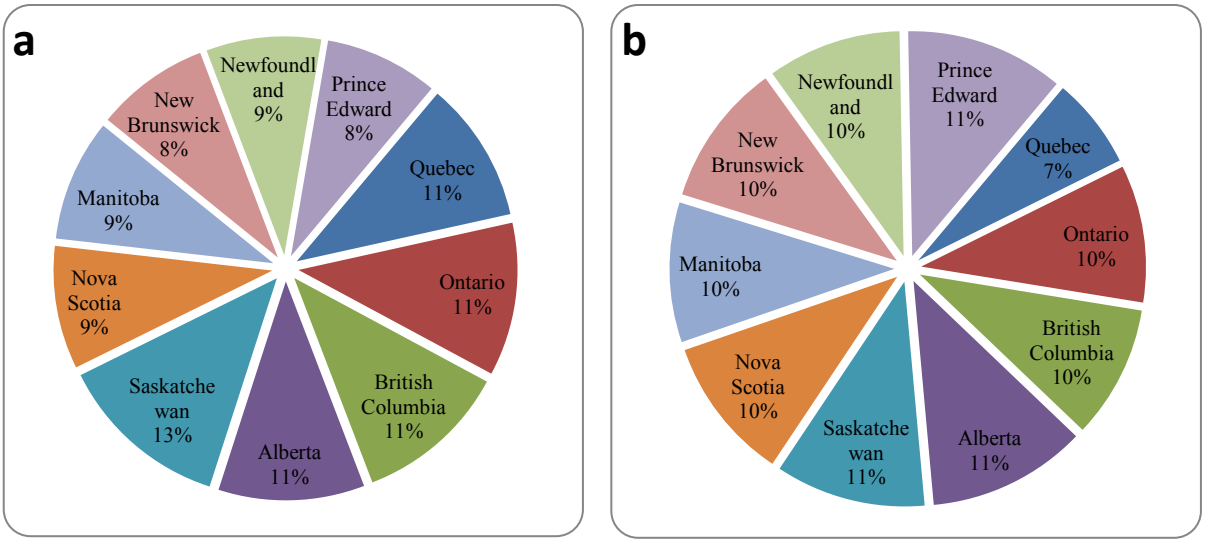


Figure 6. a) Average share of total funding per researchers in Canadian provinces, 1996-2010, b) Average share of total number of articles per researchers in Canadian provinces, 1996-2010

Figure 6-b shows the average provincial share of total number of articles for the NSERC funded researchers. Almost all the Canadian provinces have the same share of the total number of articles (10-11%) except the researchers from Quebec (7%). More interestingly, when we compare the results from the Figures 6-a, and 6-b, it can be seen that although Quebecers have relatively high share of the total funding the average number of articles that they have produced is the lowest. This is a preliminary finding and we will further investigate other important factors, like the quality of the papers. We will now take the number of researchers into account to investigate and compare the average funding available to the researchers in the Canadian provinces.

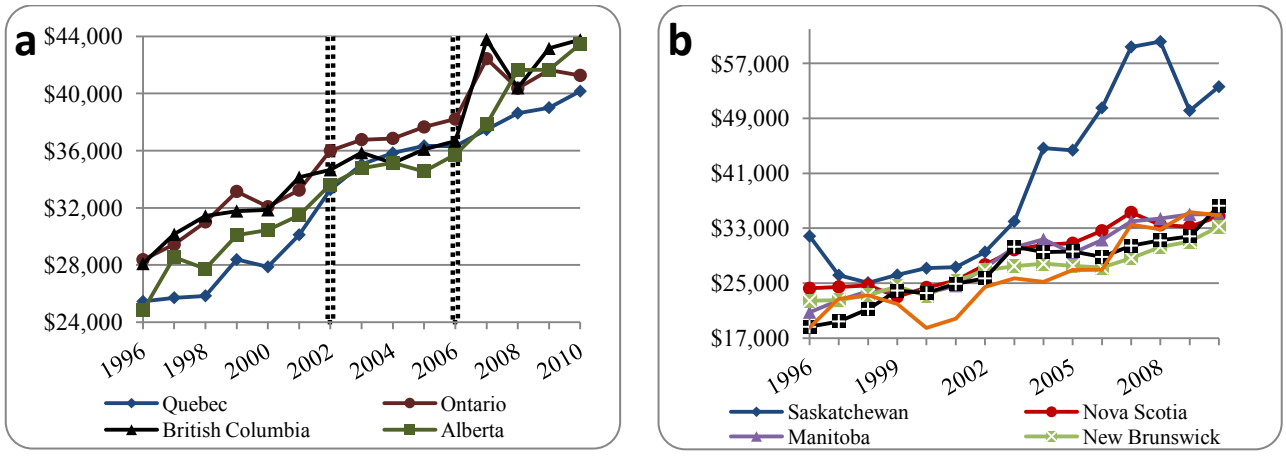


Figure 7. a) Funding per researcher in the high funding provinces, 1996-2010, b) Funding per researcher in the low funding provinces, 1996-2010

According to Figure 7-a, average funding trends of the researchers of high funding provinces can be divided into three periods as follows:

- *Period I*, (1996 to 2002): high annual increase of funding
- *Period II*, (2002 to 2006): low annual increase of funding
- *Period III*, (2006 to 2010): high annual increase of funding

We will refer to these funding periods in the rest of the paper as Period I, II, and III. As it can be seen in Figure 7-a, apart from Period II where the average amount of funding per researcher is only slightly increasing for most of the years, in the other parts of the time interval we see a more increasing trend. However, researchers from Quebec are receiving lower amount of money almost during the whole period. This difference is most obvious during the first and last three years of the time interval.

Figure 7-b shows the same indicator for the provinces in low funding provinces. Except Saskatchewan, a slight annual increase with almost similar slope is observed in the other provinces throughout the time. The increase is more notable for the researchers of Saskatchewan especially after 2002 where their trend completely departs from the others. More interestingly, after 2003 the average amount of funding for the researchers of Saskatchewan becomes considerably higher, reaching even much above the levels of the researchers from the high funding provinces. This jump might be due to the NSERC support of the research facilities in Saskatchewan. During the mentioned period, a considerable amount of money (about $63 Millions in total) has been allocated to Saskatchewan through the "*Major Facilities Access*" program[7]. The aim of this grant is to support academic research institutes, resources, and facilities (NSERC, 2012b). In the next part we evaluate the impact of funding on the number of publications.

5.1.1.3.2 Funding and Rate of Publications

Apart from the total amount of articles and funding allocated, it could be informative if we consider the trends of the mentioned factors during the examined time interval. According to Figure 8-a, funding has had an increasing trend during almost all the years where it reached to its maximum in 2010 for all the four provinces. However, Ontario has

[7] Was replaced by the "*Major Resources Support*" program in 2006.

received significantly more money than other provinces in our first group and is also producing more articles respectively. More interestingly, the trend of articles can be divided into three different periods. From 1996 to 2001 and from 2007 to 2010 the number of articles has remained almost constant for all the four provinces under study. The constant trend in the number of articles in the mentioned periods is quite interesting since it is not in line with the increasing amount of funding in the respective time intervals. Moreover, from 2001 to 2007 we see a drastic increase in the number of articles in all the provinces. There is a possibility that researchers focused more on other factors (*e.g.* quality of the papers) rather than the quantity of the articles during the constant periods. Moreover, from figure 8-a it can be said that the curves for funding and articles for Ontario, British Columbia, and Alberta are closer to each other in comparison with Quebec. This is in line with our findings from Figure 6 that indicates the share of article production for the researchers from Quebec is lower than their share from total funding.

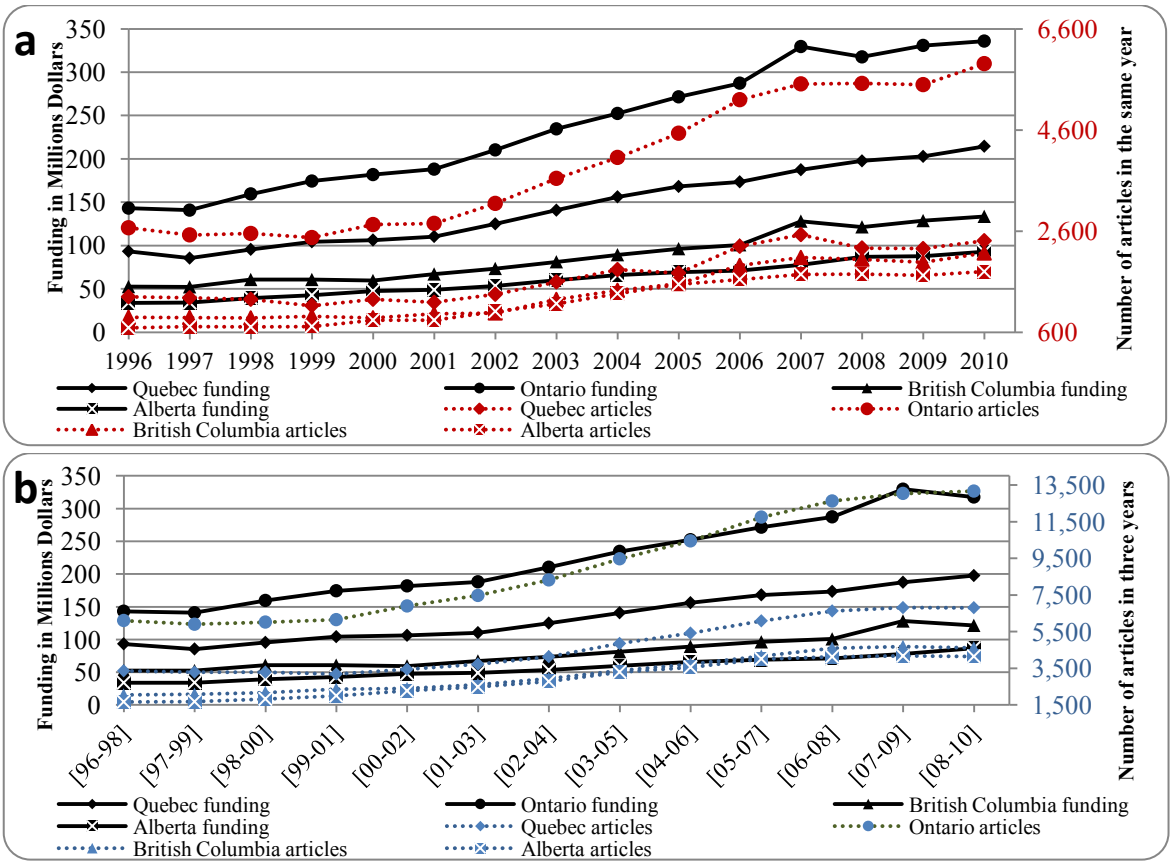


Figure 8. a) Publication rate and funding in high funding provinces, 1996-2010, b) Publication rate in a three-year time window from the period of [1996-1998] to [2008-2010] and funding from 1996 to 2008 in high funding provinces

In addition, due to the fact that funding needs three to five years to show effect, we defined a three-year time window for the number of publications for each of the funding years. As an example, we gathered all the articles for the period of 1996 to 1998 that were published by the researchers who received funding in 1996. According to Figure 8-b, trend of articles is almost the same as the trend of funding. In other words, whenever we see an increase in funding the number of publications of the funded researchers has been also augmented, and vice versa. Moreover, a significant increase in number of publications is observed from 2000 to 2006 for the researchers of high funding provinces which is almost the same as Figure 8-a. Although the funding is following an increasing trend in almost all the years, the trend of number of publications has become almost constant during the last three examined periods. This could be due to the fact that the raise in total funding has been concurrent with a higher increase in the number of funded researchers in a way that the average amount of funding decreased.

The trend of number of articles in the same year of funding for low funding provinces (Figure 9-a) is following the same trend as the one for the high funding provinces except for Prince Edward province where the amount of funding and number of articles is much less than the others that makes its trend looks more constant during the whole time interval. In addition, the amount of funding for the provinces of the low funding group has not always increased, especially for Saskatchewan and Nova Scotia where we see a considerable drop in the amount of funding after 2007. This may acknowledge higher attention of NSERC to the high funding group of provinces, probably because most of the high ranking universities and research institutes are located in that group.

Figure 9-b shows the results of the same analysis considering a three-year time window for the publications. Unlike Figure 9-a, in Figure 9-b a smooth increasing trend is observed in the number of articles except for the last three periods (after the period of [2006-2008]) where a constant or slightly decreasing trend is seen. This drop is almost in line with the trend of funding in the examined provinces. Hence, it seems that there is a direct positive relation between total funding and number of publications of the funded researchers in the low funding group of provinces.

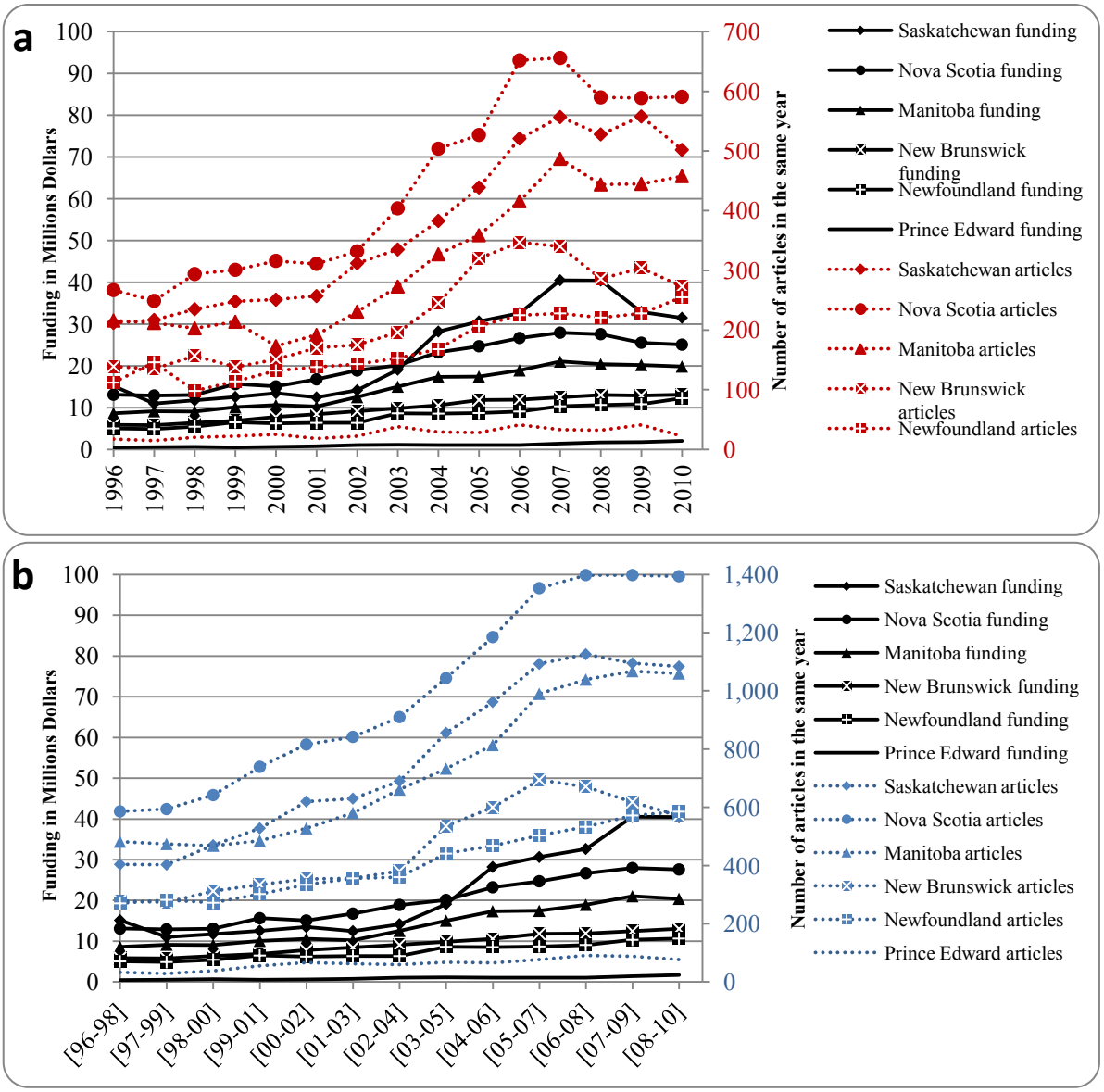


Figure 9. a) Publication rate and funding in low funding provinces, 1996-2010, b) Publication rate in a three-year time window from the period of [1996-1998] to [2008-2010] and funding trend from 1996 to 2008 in low funding provinces

To further investigate the relation between funding and publication, we take into consideration the average number of articles produced per researcher in a three-year time window. Dashed vertical lines in Figure 10-a represent the funding periods that were defined earlier. We did not include Prince Edward in Figure 10-b since the trend of the average articles per researcher was very sinusoidal with considerable differences between maximums and minimums. This was quite predictable since the total amount of funding allocated to the researchers of Prince Edward, the total number of publications, and also the number of researchers there is much lower than in the other provinces. Comparing the trends in

Figure 10-a with the funding periods, it can be said that no significant impact of funding is observed on the average number of publications produced in a 3-year time interval in high funding group of provinces since the trend of the article production (Figure 10) does not follow the funding trend (Figure 7).

Another point observed from Figure 10-a and Figure 10-b is that the productivity of researchers (in terms of the average number of publications per researcher) in both low and high funding groups of provinces is quite comparable. This is interesting since the average funding allocated to the high funding provinces is much higher than what was allocated to the ones in the low funding group, except for Saskatchewan that was discussed earlier. Moreover, according to Figure 10-a researchers from Quebec are showing very low productivity. One of the possible reasons for such a low productivity could be the language factor in a way that there is a possibility that the works of French speaking researchers were less counted in our analysis since Scopus is English-biased and non-English articles may be underrepresented. Finally, from Figure 10 it seems that after the period of [2006-2008] an almost non-increasing trend (even declining in some cases) is observed in the amount of average number of articles per researcher in both low and high funding groups of provinces which is concurrent with the high increase in the Period III of funding.

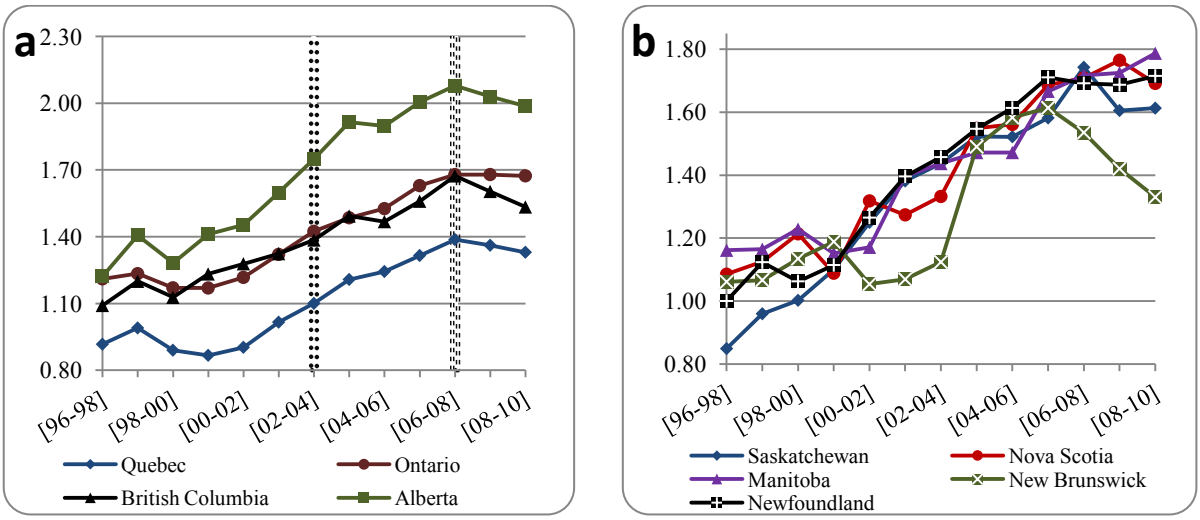


Figure 10. a) Number of articles produced per researcher in 3- year time window in high funding group of provinces, b) Number of articles produced in per researcher in the 3- year time window in low funding group of provinces

We also analyzed cost of the papers in Canadian provinces measured by "*Dollar Per Article*" (DPA) indicator that is defined as the ratio of funding to the number of articles produced in a three-year time window. According to Figure 11-a, the curves of cost of the papers have a positive slope for the high funding group of provinces from [1996-1998] to [2000-2002] and from [2006-2008] till [2008-2010]. However, it decreases from the period of [2000-2002] till [2006-2008]. The funding periods that were defined earlier are represented by dashed vertical lines in Figure 11-a. As it can be seen in Figure 10-a, during Period II of funding when the rate of increase in the amount of funding is lower the rate of publication does not decline and they increasingly produce articles hence the price of articles (DPA in Figure 11-a) goes down. However, when the rate of increase in the amount of funding is higher in Period III of funding the rate of publication decreases (Figure 10-a) while DPA starts to increase (Figure 11-a) indicating higher price of article production. Therefore, our findings from Figure 11-a confirm the previous results from Figure 10-a that there is no significant impact between funding and article production in high funding group of provinces.

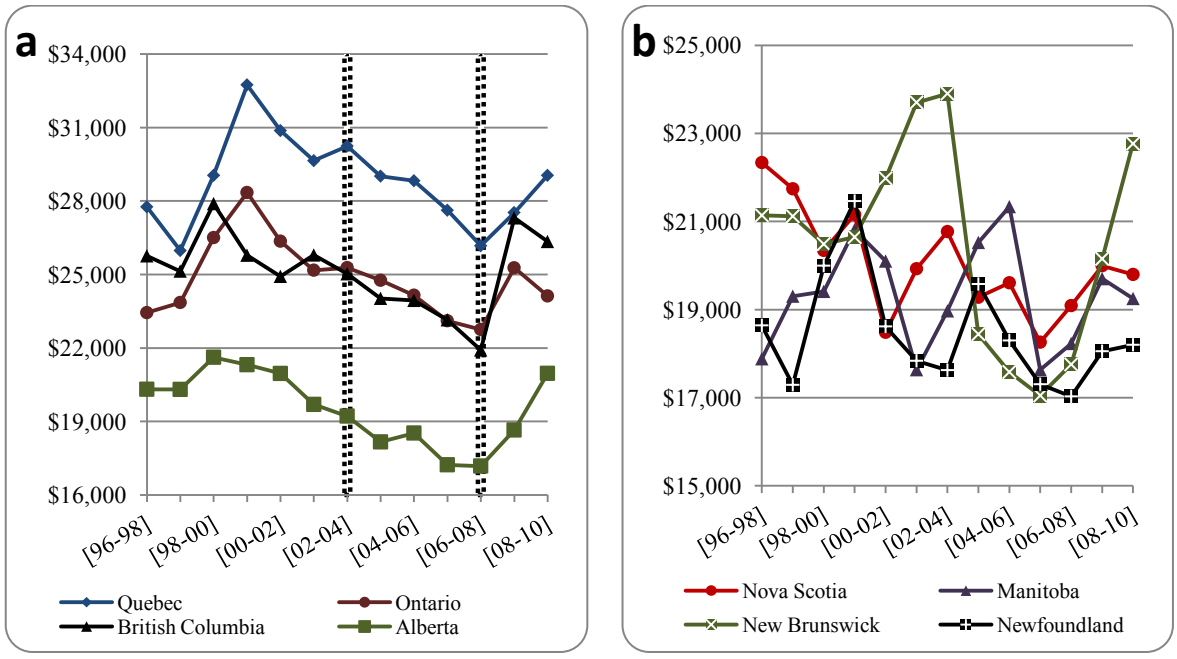


Figure 11. a) DPA in high funding provinces in 3-year time window, [1996-1998] to [2008-2010], b) DPA in low funding provinces in 3-year time window, [1996-1998] to [2008-2010]

DPA curves for Saskatchewan and Prince Edward provinces were considered as outliers and excluded from Figure 11-b due to the reasons that were mentioned earlier (*i.e.* the

numbers were small for Prince Edward province in comparison with other provinces and the considerable increase in NSERC funding for Saskatchewan researchers). For the remaining provinces of the low funding group in Figure 11-b, a quite constant trend without any severe fluctuation is observed (except for New Brunswick where some fluctuations is seen in the middle periods). Therefore, it can be said that funding has a positive impact on publication rate in the low funding group of provinces that supports our previous finding from Figure 10-b.

In other words, funding follows a slightly increasing trend in low funding provinces during the whole examined period (Figure 7-b) and the average rate of publication is increasing as well (Figure 10-b). As we see an almost constant trend of price per article in Figure 11-b, it can be concluded that there might be a positive impact between funding and output. One reason of the bigger impact of funding on the publications rate in low funding provinces in comparison with the high funding group could be that the researchers in low funding provinces might use the available money more efficiently since the amount is limited. However, in rich provinces researchers might not care too much about using money efficiently as the in hand money could be more than what they really require. In the next section we investigate the quality of the publications.

5.1.1.3.3 Funding and Publications' Quality

In this section the impact of funding on the quality of publications is investigated. We considered two proxies for quality of the papers, one is based on the citation counts and the other one is based on the impact factor of the journals in which the articles are published. Both of them can serve as a proxy for quality, but with a slightly different meaning. Impact factor indicates the respectability of the journal, *i.e.* quality and level of contribution perceived by the authors and the reviewers of the paper, whereas citations show the impact of the article on the scientific community and on the subsequent research. Since both proxies have some flaws, we decided to include both of them.

Figure 12 depicts the trend of average number of citations in a three-year time interval. For calculating the average number of citations we took the year of funding into consideration. As an example, for calculating the average number of citations in the period of [1996-1998] we first gathered all the articles for the period of 1996 to 1998 for all the

researchers who received funding in 1996. As the next step, we defined a three-year time interval for each article and count its number of citations. Therefore, citations were counted within the period of 1996 to 1998 for the articles published in 1996 where for the articles published in 1998 citations were counted from 1998 to 2000. Finally, we averaged the number of total citations over the number of articles. As it can be seen, the trend of average number of citations is almost the same for the both groups of provinces. However, Figure 12-a nicely follows the trend of funding periods (defined in Figure 7). In other words, whenever we see a significant raise of funding in Figure 7 (Periods I and III) a considerable increase is also observed in Figure 12-a. Hence, funding seems to have a significant impact on the quality of the papers in high funding provinces.

As explained earlier, funding follows a slightly increasing trend in low funding provinces. However according to Figure 12-b, the increase in the quality of the papers is seen till the period of [2003-2005] while after that we see some fluctuations in all the provinces of the low funding group. Hence, it seems that there is no relation between funding and the quality of the papers in low funding provinces especially after the period of [2003-2005]. Moreover, as expected, the average number of citations is higher for the researchers located in the high funding group of provinces (Figure 12-a) in comparison with their counterparts in the low funding provinces (Figure 12-b).

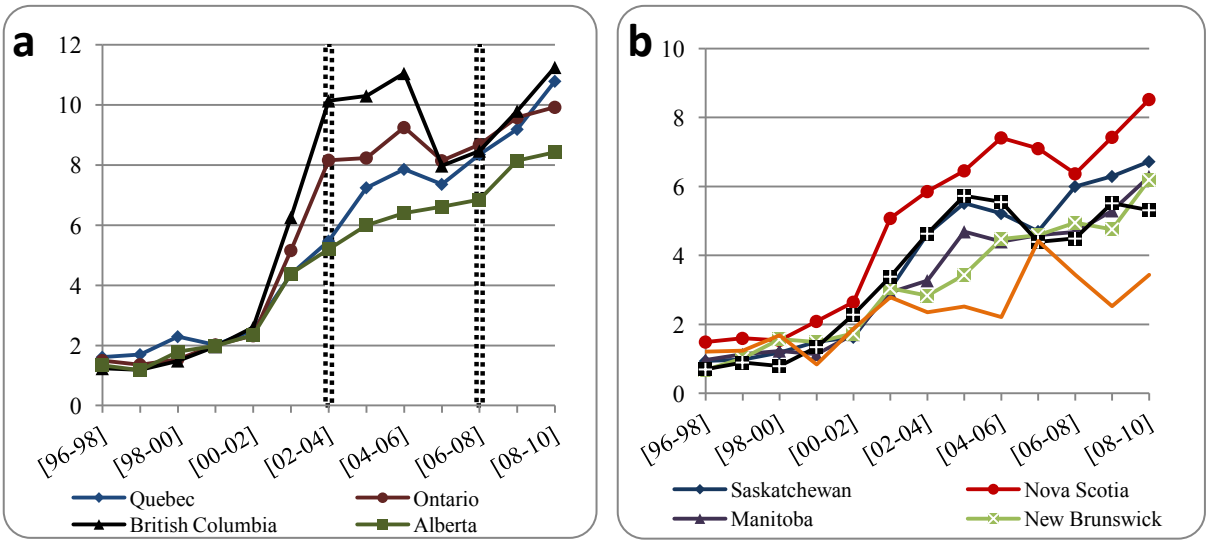


Figure 12. a) Average citation counts in high funding provinces, [1996-1998] to [2008-2010], b) Average citation counts in low funding provinces, [1996-1998] to [2008-2010]

Figures 13-a, and 13-b depict quality of the papers in a three-year time window measured by the average impact factor of the journals in which the articles were published. As it can be seen, based on the calculated measure the average quality of the papers published by the researchers in the high funding group of provinces is higher than the low funding provinces. Level of the calculated impact factor proxy reaches between 4 and 5 in the last years in the high funding group of provinces while the value is only around 2.4 to 4 for the low funding group of provinces. Clearly, researchers from the high funding provinces on average publish in higher quality journals than their counterparts in the low funding group. One reason could be a higher number of high ranking universities in the high funding group of provinces in comparison with the low funding group since it is more probable for researchers in the high ranking universities to publish on average in higher ranking journals. In addition, higher average money available in the high funding group may enable them to improve the quality of their work through different ways like supplying more modern equipments, employing more skillful experts in their research teams, forming larger research teams, *etc*. Hence, these better conditions might enable them to do a higher quality research which can get published in a higher quality journal.

According to Figure 13-a, although researchers from Alberta have shown a considerable progress in the quality of their work in the last year, papers of the researchers from Quebec and Ontario have had the highest quality. Apart from the language factor that was already discussed, high quality articles can also justify our findings from Figure 10 where researchers from Quebec had the lowest average productivity. In other words, it seems that researchers in Quebec focus more on the quality of their work rather than the quantity, by publishing in higher quality journals. In addition, although there are some gaps between the curves for different provinces, they get close to each other as we reach the latest periods. Hence, the overall trend for the high funding provinces indicates that researchers are tending to higher quality work. As it can be in Figure 13-b, it seems that this issue is less important for the researchers of the second group of provinces although a slight increase in the level of average quality can be observed.

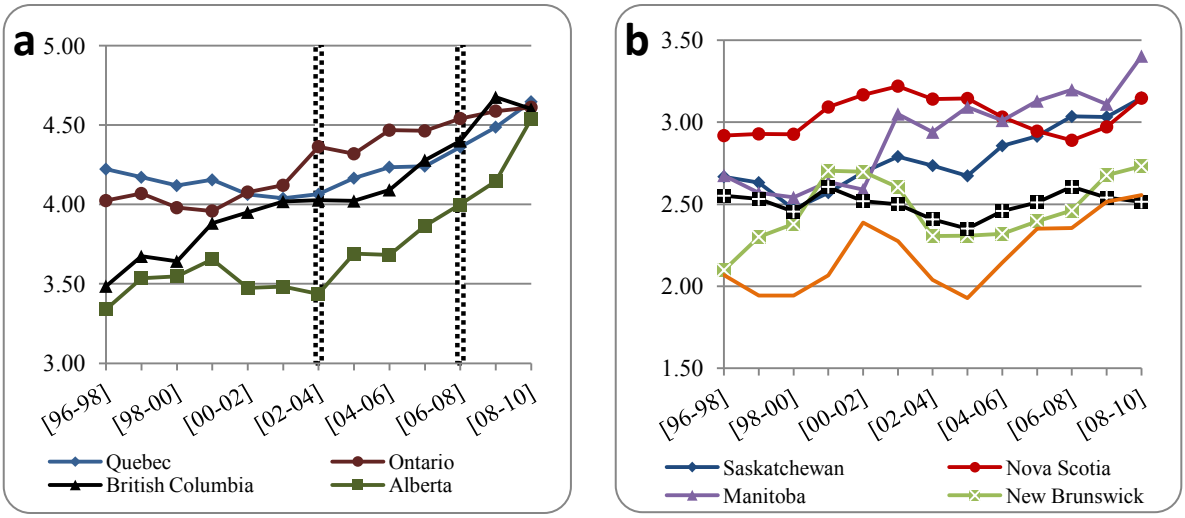


Figure 13. a) Average journal impact factor in high funding provinces in 3-year time window, b) Average journal impact factor in low funding provinces in 3-year time window

Comparing our findings from Figure 13 with the funding periods indicated in Figure 7, it can be said that no impact of funding is observed on quality of the works of the researchers in the high funding provinces as the journal impact curves in Figure 13 do not follow the funding trends. For example during Period I when researchers get relatively high level of funding the trend of the publication quality of all the provinces in high funding group is almost constant except for British Columbia. For the low funding group of provinces, a positive relation between funding and quality of the works can be seen only for the researchers located in Manitoba and Saskatchewan where the slowly increasing line in fact corresponds to the increasing trend of funding in Figure 7-b. However, for the other provinces of the mentioned group no impact is observed and comparing this finding with our results from Figure 10-b, it seems that researchers located in the mentioned provinces have more focused on increasing the number of articles produced, paying less attention to the quality of the work.

In order to evaluate the impact of funding on quality of the papers more accurately, we defined "*Dollar Per Impact Factor*" (DPIF) indicator as follows:

$$DPIF_{[i,j]} = {avg(funding)_{year=i}} \Big/ {avg(impact\ factor)_{year\ \in [i,j]}}$$

In other words, DPIF implicitly shows the amount of average money invested on quality of the papers in a sense that higher DPIF for a group of researchers means that those

researchers' articles have on average lower impact factor per dollar invested, and vice versa. Figures 14-a and 14-b depict the results for the Canadian provinces. According to Figure 14-a, the curves follow an increasing trend till the period of [2002-2004] after which they slightly decrease. In other words, after continuous increase for seven periods, the average cost of quality of the papers for the researchers of the high funding group of provinces gradually decreases. Interestingly, this cost is the lowest for the researchers of Quebec during almost the whole time interval. From Figure 13-a it was observed that average impact factor of the journals follows a slight increasing trend. On the other hand the trends in Figure 14-a follow the similar trends of funding periods except for Period III. Hence, this partially confirms our previous finding that it seems there is no impact of funding on the quality of works in the high funding group of provinces.

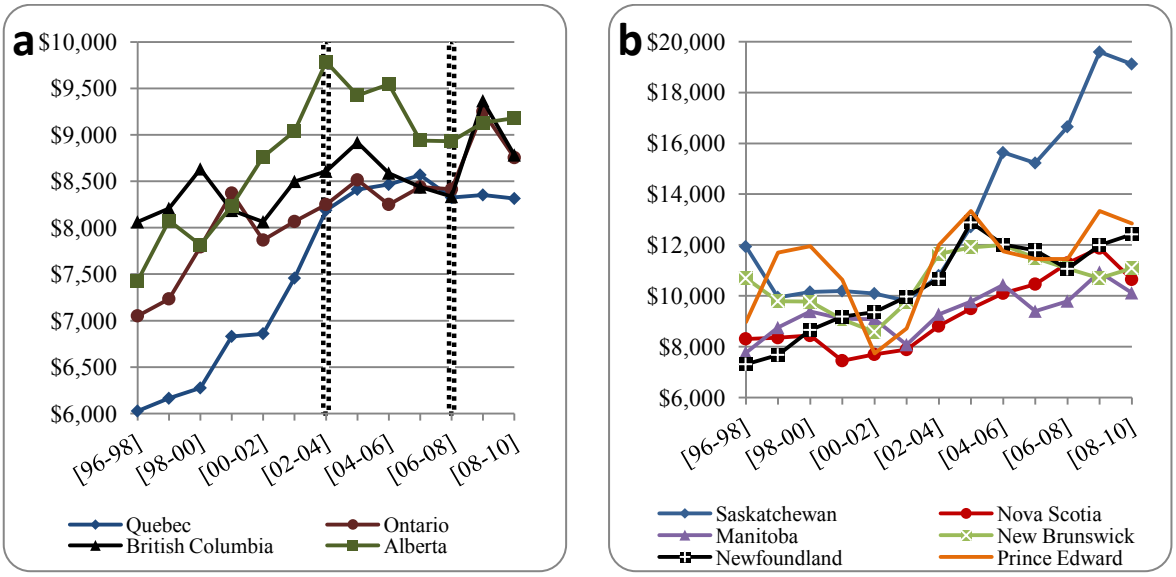


Figure 14. a) DPIF in high funding provinces, [1996-1998] to [2008-2010], b) DPIF in low funding provinces, [1996-1998] to [2008-2010]

As it can be seen in Figure 14-b, there is no general trend over the whole examined period for all the low funding provinces and a lot of fluctuations is observed. The case is more severe for Saskatchewan and Prince Edward provinces. For Saskatchewan a significant increase is seen after the period of [2003-2005] possibly due to the special support of NSERC allocated to the researchers that was already discussed. In general, as the curves of Figure 14-b do not follow the slight increasing trend of funding in low funding provinces (Figure 7-b) no conclusion can be made about the impact of funding on the quality of the works. This is also in line with our findings from Figure 13-b. In addition, the levels of DPIF

for the low funding provinces are on average higher than the ones in the high funding provinces that indicates relatively higher cost of quality in low funding provinces.

Our findings from Figures 12, 13, and 14 indicate that the change in funding does not seem to have much impact on the quality of journals the researchers publish their articles in while it affects the average citation received by articles especially for the researchers located in the high funding provinces. Hence, it can be said that even if researchers receive lower amounts of funding they do not lose their ambitions and still continue submitting their papers to the high ranking journals. Also, as they might be already well-known and respected in the scientific community their articles would get accepted in high ranking journals thanks to their reputation. However, citations received by their articles shows the real impact and quality of their work better where we see that in the years of low annual increase of funding (Period II) citation levels were basically constant (decreasing in some cases). Therefore, it seems that funding does affect ability of the researchers to create a highly cited work. In the next section, we evaluate the impact of funding on the scientific collaboration of the researchers.

5.1.1.3.4 Funding and Collaboration

Collaboration can generate large advantages for the society. Through collaborative scientific activities, different skills and ideas are combined and resources are thus used more efficiently. This can bring economies of scale in scientific activities and may avoid research duplication (Ubfal & Maffioli, 2011). In addition, collaboration trains the available skills that will result in development of new expertise (Lee & Bozeman, 2005). Funding can influence the collaboration patterns among the researchers. Higher level of funding can enable funded researchers to expand their scientific activities that may result in higher scientific production. A great advantage of funding is that it enables researchers to cover the collaboration costs (*e.g.* finding right partners and research coordination). Moreover, it allows the central researchers[8] to better internalize some of the required duties through the coordination (Ubfal & Maffioli, 2011). According to Porac *et al.* (2004), availability of funding may help the central scientist(s) to make a balance between new knowledge creation

[8] More central researchers have more connections and tend to have favored positions in the collaboration network.

and management of the existing collaborative relationships in the collaboration network. On the other hand, one may notice that higher amount of funding is not always beneficial for the network. If the collaboration network is at its social optimum level then allocating more funding can affect the system negatively by adding more collaboration links (Ubfal & Maffioli, 2011).

However, measuring the scientific collaboration is not easy. In the literature, co-authorship is known as the standard way of measuring collaboration since it is considered as a sign of mutual scientific activity (De Solla Price, 1963; Ubfal & Maffioli, 2011). Co-authorship as a measure is practical, invariant, verifiable, inexpensive (Subramanyam, 1983), and quantifiable (Katz & Martin, 1997). To analyze the co-authorship patterns of the researchers, we use average number of "*Authors Per Article*" (APA) as a proxy of the team size. The results are depicted in Figures 15-a, and 15-b. The funding periods are shown by dashed vertical lines in Figure 15-a.

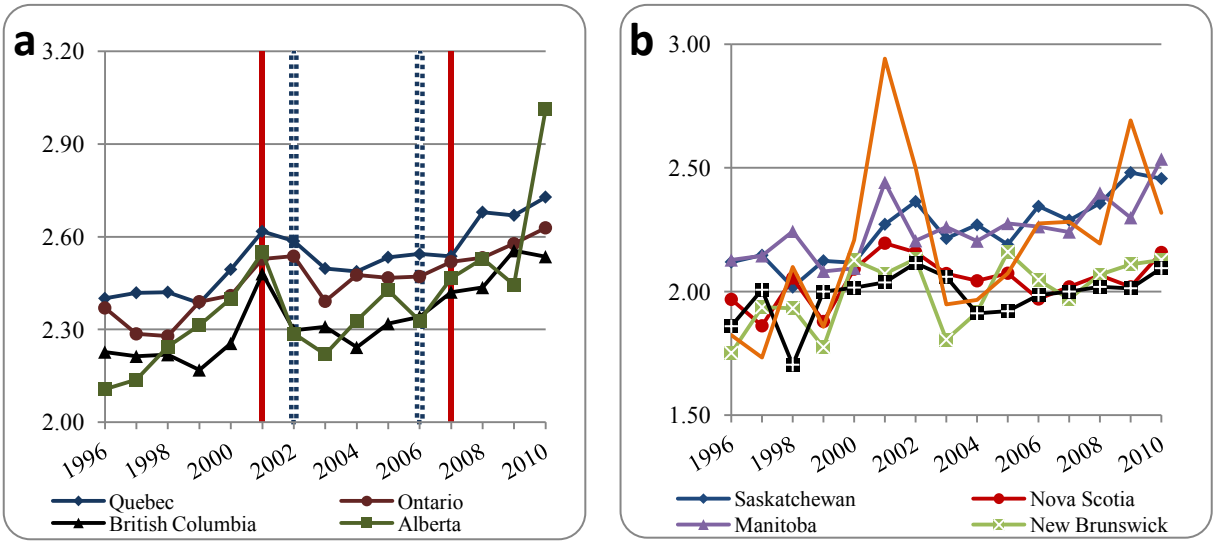


Figure 15. a) Authors per article (APA) in high funding provinces, 1996-2010, b) Authors per article (APA) in low funding provinces, 1996-2010

As it can be seen, we can divide the trend of APA in the high funding provinces into three periods as indicated by vertical red lines in Figure 15-a. For the high funding provinces, APA follows an increasing trend from 1996 to 2001 and from 2007 till 2010 where from 2001 to 2007 we see a slightly decreasing trend. Interestingly, almost similar trends are observed for the funding amount of the researchers of the high funding group of provinces (dashed vertical lines). Hence, it can be said that in the periods with a high annual

increase of funding the teams of the funded researchers also became larger. Comparing the slightly increasing trend of funding for the low funding provinces (Figure 7-b) with Figure 15-b, we can state that the same proposition can be also valid for researchers of the low funding group of provinces as their trend of APA is also augmenting moderately. Another interesting point is that in general the average team size in the high funding provinces, which ranges between 2.5 to 3 researchers in the latest periods, is larger than in the low funding group of provinces, where on average a team has only around 2 to 2.5 members. This is also expected since researchers located in the high funding provinces benefit from higher average level of funding that might help them to better expand their team sizes and scientific activities.

5.1.1.4 Conclusion

Stunning progress in information technology and the availability of more accurate and integrated data in one hand and the considerable amounts of annual investments on R&D on the other hand, has encouraged data scientists to focus more on the scientific evaluations. Several factors can influence scientific activities where financial support and collaboration patterns are among the most important ones. According to our results, funding seems to have played an important role not only in enhancing scientific productivity of the researchers but also in the formation of scientific teams and collaboration patterns.

In this research we divided the Canadian provinces into two separate groups namely, high and low funding provinces. Almost all the Canadian provinces had quite comparable total funding shares per researcher. In addition, the total productivity of the researchers from all the provinces was almost at the same level. However, as it was observed funding shows different effects in the mentioned groups of provinces. Although the increase in the amount of average funding has been followed by higher rate of publication in the low funding provinces, we found no impact of funding on number of papers in the high funding group. This can be due to the fact that researchers who reside in the low funding provinces might use their available funding more effectively and more efficiently while in the high funding provinces researchers might allocate their extra available funding to the activities that will not necessarily result in higher number of publications. Therefore, it seems that impact of funding on the scientific production follows an inverted U-shaped curve. That means higher

funding results in higher number of publications (Arora & Gambardella, 1998; Zucker, *et al.*, 2007) but after a certain level (when the researcher is overly rich) the effect of funding on the output decreases.

On the other hand, our results suggest a positive relation between funding and quality of the works (measured by citations) in the high funding group of provinces while it does not affect the quality in the low funding provinces. It was quite expected since higher ranking universities and research institutes are mostly located in the high funding provinces. In addition, researchers who are working on high priority research projects are mainly located in the high funding group of provinces. For example in 2010, 1,023 researchers in the high funding provinces received funding through NSERC strategic programs (that focus on high priority projects) that is about 8.5 times higher than the ones who were located in the low funding provinces (123 researchers). Hence, apart from other potential influencing factors (*e.g.* research policies, cultural issues) working on sensitive high priority projects and working in better-established scientific teams in higher ranking universities and institutes might be some of the reasons of the higher quality of works in the high funding universities. The positive impact of funding on the scientific output has been also confirmed in the study of Godin (2003) who assessed the impact of NSERC funding in the period of 1990-1999. However, he found no impact of funding on the quality of the papers where in our study a positive relation was observed.

We also used the average journal impact factor as another proxy for quality of the papers. According to the results, in the high funding group of provinces no impact of funding is observed on the average impact of the journals in which researchers have published their articles. Moreover, in the low funding group of provinces also no relation is seen except for Manitoba and Saskatchewan. This is quite interesting that the results are different for the two quality measures. It can be said that researchers might benefit from higher amount of money available to produce a paper that will be highly acknowledged in their internal scientific community by being highly cited. However, higher amount of funding does not influence their decision to aim for publishing in higher quality journals.

Regarding the co-authorship patterns, in general the trend showed the interest of the researchers to be increasingly more involved in larger research teams. In addition, a positive

impact was observed for funding on the team size both in high and low funding groups of provinces. Hence, it seems that higher amounts of funding available enable researchers to expand their scientific activities by forming larger teams and getting involved in larger projects. This partially confirms the importance of the role of funding in the formation of scientific collaboration patterns.

5.1.1.5 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, non-English articles are underrepresented (Okubo, 1997). Secondly, since Scopus data is less complete before 1996, we had to limit our analysis to the time interval of 1996 to 2010. Another inevitable limitation related to 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. In collecting articles of the funded researchers we assumed that all the funded researchers acknowledge the support of NSERC in the article. This assumption is based on the fact that according to NSERC guidelines funding has to be acknowledged in the articles of funded researchers. However, it is also probable that some researchers do not acknowledge the source of funding in their papers.

Different scientific disciplines follow different patterns in publishing articles, collaborating with other researchers, or even getting and allocating grants to the tasks. Hence to better examine scientific productivity and efficiency, a future work direction could be assessing the impact of funding on the rate of publications for different scientific disciplines separately. In addition, the impact could be separately analyzed for different types and programs of funding, and also other funding councils can be considered as the source of funding data. The analyses comparing the efficiency of different funding organizations may help the decision makers to set the best funding allocation strategy.

We used co-authorship as a measure of scientific collaboration. Other proxies (*e.g.* sub-authorship) can be also used to evaluate the cooperation among researchers. In addition, not

necessarily the collaboration among two given researchers will result in the form of a joint article. An example can be the informal relations between researchers which cannot be detected by co-authorship measures. The final limitation is in regard with using bibliometric indicators where employing other approaches (*e.g.* statistical analyses) is suggested to assess the interrelations among the variables more accurately.

5.1.2 Investigating Scientific Activities in Various Disciplines and the Impact of Different Funding Programs

Role of funding in stimulating scientific activities has been studied in the literature. However, different funding programs may have variant influence on scientific development. Moreover, scientists from different scientific disciplines may follow diverse patterns in collaborating with other researchers, using funding resources, publishing articles, *etc*. Considering the notable amounts of funding being invested on R&D activities annually, it is essential to evaluate the performance of different funding programs and scientific disciplines. This will surely help the decision makers to set better funding allocation strategies in an aim to increase the scientific potential of the country. This section, investigates the effect of various NSERC funding programs on rate and quality of publications of the funded researchers. In addition, scientific collaboration patterns of the funded researchers is investigated for different scientific disciplines and funding programs. For this purpose, we used a 3-year panel data of journal publications of the NSERC funded researchers from [1996-1998] to [2008-2010], and the amount of funding allocated to them during the period of 1996 to 2008. Bibliometric analysis was used to assess the scientific development of the funded researchers and their scientific collaboration patterns. According to the results, patterns of co-authorship and article publication highly differ in various scientific disciplines. In addition, high-priority and more specific funding programs not only resulted in higher rate of article production but also quality of the published papers was higher.

5.1.2.1 Introduction

Although evaluation of the scientific research backs to about sixty years ago (Godin, 2002), it was over the last thirty years that scientists showed an increased interest to the different aspects and effects of such an evaluation. The economic depressions and average

R&D budget cuts[9] during the past years have obliged decision makers to better allocate their available funding while on the other side researchers have been forced to justify the importance and priority of their work more than before. Hence, apart from the effects of funding on research productivity (Martin, 2003; McAllister & Narin, 1983), the level of research funding and the funding allocation procedures are playing a crucial role for encouraging the scientific development.

According to De Solla Price (1965), there are differences in the output type of researchers' scientific activities. In other words, scientists tend to publish their results as scientific articles whereas technologists do not normally publish papers. Moreover, different scientific disciplines may follow various patterns of collaboration and funding allocation, and as a result may have differences in the rate of publications. As an example, since engineers are also involved in other activities (*e.g.* engineering design), we may expect a lower productivity of them in terms of the number of publications (Gingras, 1996). Or, in humanities most of the papers are single-authored while in engineering most of the papers have more than one author. Hence, studying the impact of funding in different scientific disciplines can be informative. Moreover, the considerable amount of funding is usually allocated to a team of researchers rather than a single one (Beaudry & Allaoui, 2012), especially in high-priority university-industry research projects. Therefore, studying the trend of collaboration among the funded researchers can better reveal the effect of funding.

Several studies assessed the impact of funding on scientific productivity (*e.g.* Boyack & Borner, 2003; Payne & Siow, 2003; Jacob & Lefgren, 2007). In addition, analyzing the impact of different sources and types of funding has also attracted the attention of the researchers. A number of studies focused on the effect of contractual funding on the quantity and quality of the scientific publications (*e.g.* Arora & Gambardella, 1998; Carayol & Matt, 2006). Using statistical analysis, various studies investigated the impact of federal funding (*e.g.* Payne & Siow, 2003; Huffman & Evenson, 2005), industry funding (*e.g.* Gulbrandsen & Smeby, 2005), or private funding (*e.g.* Beaudry & Allaoui, 2012) on scientific productivity and research performance. Analyzing the impact of various sources of funding

[9] Although the trend of funding was sometimes steady, the number of applicants has been always increasing.

and different funding programs can help to better distinguish the most effective R&D investment organization or program that may result in better allocation of the available money.

NSERC, established in 1978, is a Canadian government agency that provides funds for scientific research through different funding programs. NSERC supports about 23,000 university students as well as 11,000 university professors. Budget of NSERC for funding programs in 2005-2006 was $859 million (NSERC, 2012a). The target areas of the programs are quite diverse. For example, one of the NSERC funding programs is named "*Discovery Grants Program*" (previously named Research Grants Program) that funds general long term research activities of the Canadian researchers instead of individual research projects. Hence, most of the Canadian researchers, especially in natural sciences and engineering, are being funded by NSERC council. Analyzing the impact of various funding programs of NSERC can reveal the effectiveness of each of the programs in stimulating the scientific performance of the funded researchers.

This paper employs bibliometric analysis to assess the impact of different NSERC funding programs on the research performance of the funded researchers in various scientific disciplines during the period of 1996 to 2010. Our motivating questions to address are: Do patterns of the scientific activities differ in various scientific disciplines? Do all the funding programs affect the scientific activities of the funded researchers in a similar way? Are better funded disciplines more productive? How the collaboration patterns differ in different scientific disciplines? The rest of the section is organized as follows: Section 5.1.2.2 describes methodology and data that is used in this study. The empirical results and interpretations are provided in section 5.1.2.3. Section 5.1.2.4 presents the findings of this research and the limitations of this study and some directions for the future work are discussed in the last section 5.1.2.5.

5.1.2.2 Data and Methodology

Almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC (Godin, 2003). Hence, we focused on NSERC funding as the input source to the R&D activities of the Canadian researchers and collected the funding data from NSERC for the period of 1996 to 2008. This led to 66,377 distinct Canadian researchers who received funding from NSERC through 135 different funding programs during the aforementioned period. In addition, researchers were categorized into seven scientific fields based on their research activity and portfolio as follows:

1. Engineering
2. Chemistry and biology
3. Mathematics
4. Health and life sciences
5. Physics and geology
6. Animal sciences
7. Other

Scopus was selected as the output source of scientific activities of the funded researchers and articles of the NSERC funded researchers was collected for the period of 1996 to 2010. We searched for all the articles that had acknowledged NSERC funding support within the body of the article. This was a crucial step in fetching the accurate data. The articles dataset totally contained 124,722 articles and 177,449 authors that acknowledged the NSERC support in the respective article. For evaluating the quality of the papers, SCImago was selected for collecting the impact factor information of the journals in which the articles were published in and the result was integrated into another dataset. SCImago was chosen for two reasons. First, it provides the journal impact factors for each of the single years of our examined time interval. This enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our articles database.

Bibliometric analysis is used to investigate the impact of various NSERC funding programs on the rate of publication and quality of work of the funded researchers as well as their collaboration patterns. In addition, scientific development in different scientific disciplines is investigated. In the literature, three-year (*e.g.* Payne & Siow, 2003) or five year (*e.g.* Jacob & Lefgren, 2007) time windows have been considered for funding to take effect. In this paper, we assume a three-year window for the funding to influence the productivity of the researchers. For example, for the funding year of 1996 we gather all the articles of the funded researchers from 1996 to 1998. This results in 13 different time intervals for article production starting from [1996-1998] to [2008-2010], where the funding period expands from 1996 to 2008.

5.1.2.3 Results

The impact of NSERC funding is analyzed from two perspectives. First, we assess the funding impact on scientific development in different disciplines. Secondly, the impacts of different NSERC funding programs are investigated and compared.

5.1.2.3.1 Impact in Scientific Disciplines

The funded researchers were categorized into different disciplines based on several factors *i.e.* their affiliation and department, the subject and area of the awarded fund, the selection committee for the researcher's grant, and the title of the grant. For this purpose, a JAVA program was coded that automatically searched for the defined criteria over the list of the funded researchers and assigned a scientific field to each of them. Hence, the funded researchers were classified into seven categories that were mentioned in 5.1.2.2. Figure 16 shows the total funding share and the total number of article production share of the aforesaid scientific disciplines from 1996 to 2010. According to Figure 16-a, as expected NSERC allocated more funding to engineering and pure sciences rather than health and life studies. Interestingly, a considerable amount of funding has been allocated to the researchers of chemistry and biology fields. Surprisingly, although the engineers have been receiving the highest amount of total funding, the total number of publication is highest for the researchers involved in chemistry and biology.

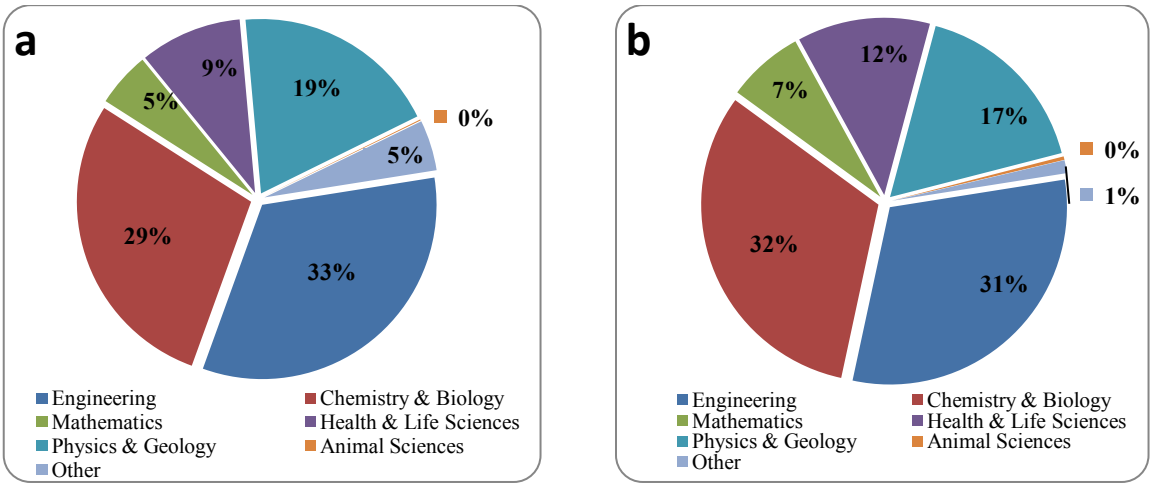


Figure 16. a) Total funding share of scientific disciplines, 1996-2008, b) Total article production share of scientific disciplines, [1996-1998] to [2008-2010]

Since the amount of funding and the number of articles produced are much lower for the fields of "*animal sciences*" and "*other*", we exclude them from the analysis and focus on the main five scientific fields in the rest of the paper. According to Figure 17-a, funding has followed an increasing trend in all the disciplines almost during the whole examined period. A drastic raise is seen after 2001 where before 2001 a steady trend is observed for some of the disciplines like mathematics. Engineers have been receiving the highest amount of money where after 2001 the gap between engineering and chemistry field (as the second highest) becomes more significant.

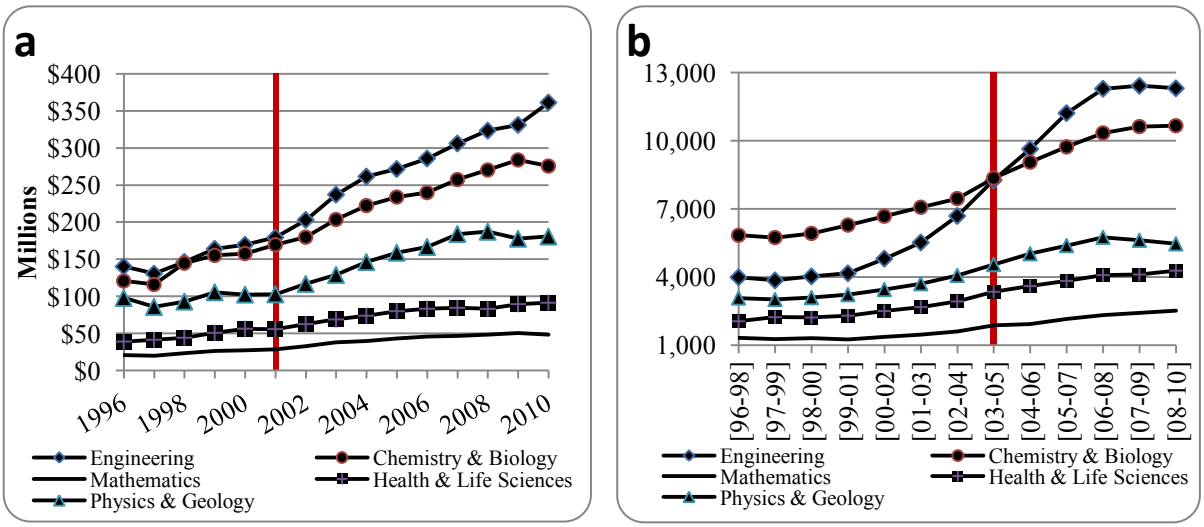


Figure 17. a) Funding trend, 1996-2010, b) Article production trend, [1996-1998] to [2008-2010]

As it can be seen in Figure 17-b, before the period of [2003-2005] chemists and biologists were producing the most number of articles while after the mentioned period engineering field took the place of chemistry as the field with the highest rate of scientific production. This might be related to the drastic raise in the amount of funding in 2001. The article production rank of the other disciplines (Figure 17-b) namely mathematics, physics, and health sciences exactly match their place in the funding diagram (Figure 17-a). Moreover, after a considerable raise from the period of [2003-2005] to [2007-2009], the trend of paper production has become steady during the last two periods. This drop is more intense for the fields of physics and geology.

Analyzing the trend of average amount of funding per researchers reveals that the physics and geology researchers have been receiving the highest amount of average funding while the mathematicians have received the lowest. The trend of average funding for the other three fields has been quite comparable. However, as it can be seen in Figure 18 researchers from the field of health and life sciences have produced on average the highest number of articles where chemists rank the second. Interestingly, engineers have had low rate of average article production per researcher in comparison with the other fields, especially before the period of [2004-2006]. One of the reasons of such a low rate of production could be the differences in the team size in various scientific fields measured by number of co-authors in an article. Since we have counted an article once for all of its co-authors, if the number of co-authors is high in a field it may results in higher rate of production. We will further investigate the productivity of the examined disciplines by taking other factors into the account.

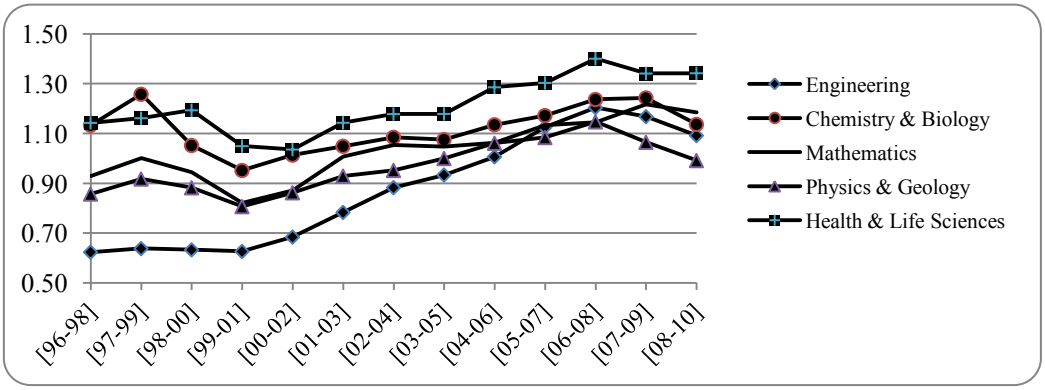


Figure 18. Average number of articles per researcher, [1996-1998] to [2008-2010]

Apart from the rate of publication, analyzing the trend of quality of the works is also important. For this purpose, we focused on the average impact factor of the journals in each of the disciplines and calculated it for all the thirteen 3-year time intervals. The results are shown in Figure 19. We cannot compare the quality of the works of different disciplines by estimating the impact factor of the journals that the articles have been published in, since the journal impact factor is highly dependent on the scientific discipline. In other words, the speed of getting cited in different scientific fields may vary that will affect the impact factor of the journals (Van Nierop, 2009). In addition, the percentage of total citations also varies among the disciplines. Hence, we just analyze the trends of impact factor in each of the disciplines separately.

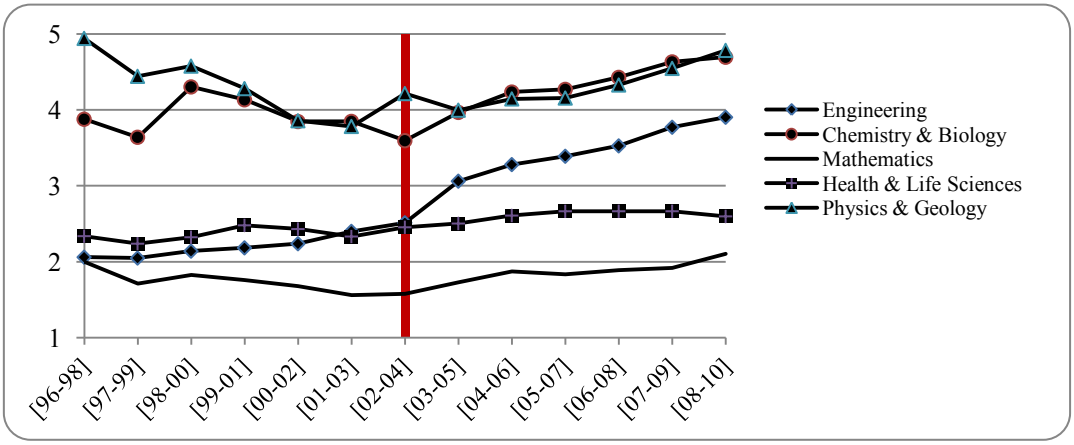


Figure 19. Average impact factor of journals in which articles were published in different disciplines, [1996-1998] to [2008-2010]

As it can be seen in Figure 19, average impact factor increases during the examined time periods in all the considered disciplines. However, a drastic raise is observed for the engineering field after the period of [2002-2004]. Comparing this finding with the results from Figure 17-a and Figure 18, it can be said that the raise in the level of funding allocated to the engineering field after 2001 could be one of the reasons that caused higher rate of article production after 2001, and higher average quality of the work of the engineers after the mentioned period. More money might have enabled them to improve the quality of their work through different ways like supplying more modern equipments, employing more skillful experts in their research teams, forming larger research teams, *etc*.

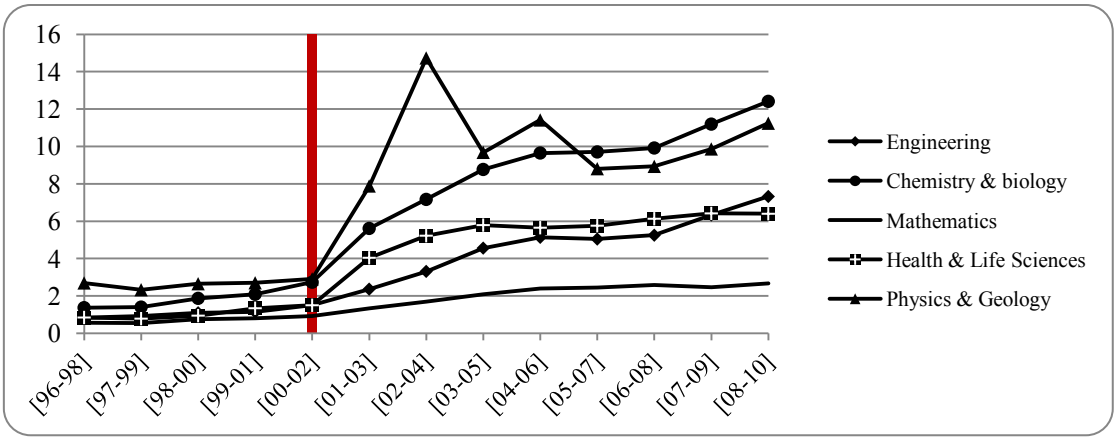


Figure 20. Average number of citations in different disciplines, [1996-1998] to [2008-2010]

We also analyzed the trend of average number of citations received by the funded researchers' articles in different scientific disciplines. According to Figure 20, we can divide the examined time interval into two periods. From the beginning till the period of [2000-2002] we see an almost steady trend in almost all the examined disciplines. However, after [2000-2002] a drastic raise is observed where the jump is the highest for the fields of physics and geology. Moreover, according to the result we can divide the examined disciplines into three parts based on their patterns of average number of citations: 1) Mathematics in which we see the lowest rate of average citation, 2) Chemistry and physics where the rate of average citations received is the highest, and 3) Engineering and health sciences that place in the middle.

To investigate the impact on scientific collaboration and research teams, we expand the analysis by focusing on the co-authorship patterns and their trends during the examined time interval. To assess the trend of multi-authorship, we used the "*collaborative coefficient*" (CC) indicator introduced by Ajiferuke *et al.* (1988). The formula is as follows:

$$CC = 1 - \frac{\sum_{i=1}^{j} \left(\frac{1}{i}\right) f_i}{n}$$

where, f_i is the number of papers that have i authors, j is the greatest number of authors, and n is the total number of articles. CC holds a value between 0 and 1 where higher value indicates higher level of multi-authorship.

Figure 21 depicts CC calculations and the trends for all the examined scientific disciplines. As expected, the level of multi-authorship is higher in chemistry, physics, and health sciences since some of the research projects in the mentioned scientific fields are performed in the research labs that may involve more researchers in a project. The mathematicians have the lowest level of collaboration in terms of multi-authorship that is also reasonable. Another observation is that after 2002 the level of multi-authorship augments drastically in all the examined fields. From Figure 17-a, it can be said that the increase in the level of funding could be one of the reasons of the augmentation in CC levels after 2002. In other words, by having more money available it seems that researchers tended gradually to expand their research teams.

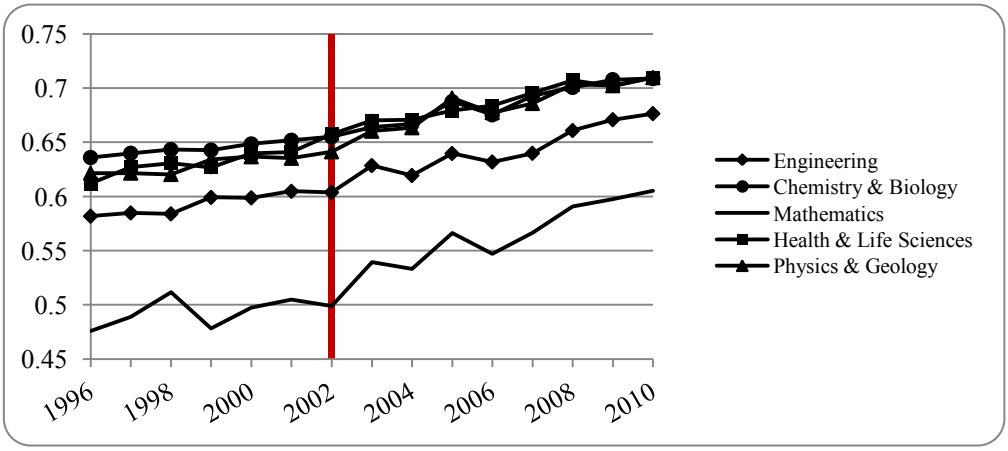


Figure 21. Collaborative Coefficient (CC), 1996-2010

To compare the scientific productivity of the researchers of the examined scientific disciplines, we introduced an indicator named $Prod_3$ that is defined for each of the 3-year time intervals as follows:

$$Prod_3[i,j] = \frac{\sum_{year=i}^{year=j} \#\ of\ papers}{\frac{\sum_{year=i} \#\ of\ researchers}{CI_3[i,j]}}$$

where, CI_3 is a modification to the "*Collaborative Index*" introduced by Lawani (1980) and is defined as follows:

$$\mathrm{CI_3[i,j]} = \sum_{\mathrm{year=i}}^{\mathrm{year=j}} \left(\frac{\sum_{\mathrm{a=1}}^{\mathrm{b}} \mathrm{a.f_a}}{\mathrm{n}} \right)$$

where, f_a is the number of papers that have a authors, b is the greatest number of authors and n is the total number of papers.

Figure 22 shows the productivity trend of the scientific disciplines during the examined time period. According to the calculated measure, we can categorize the scientific disciplines into two groups based on their productivity. Researchers from health and life sciences and mathematics show significantly higher productivity than the researchers of other examined disciplines. The main reason for such a high value is that the collaborative index for the field of mathematics is much lower than the other fields, and the number of researchers in the both fields is also lower than the others. Having less number of co-authors in a paper, and lower number of researchers resulted in higher productivity per researcher in the mentioned fields.

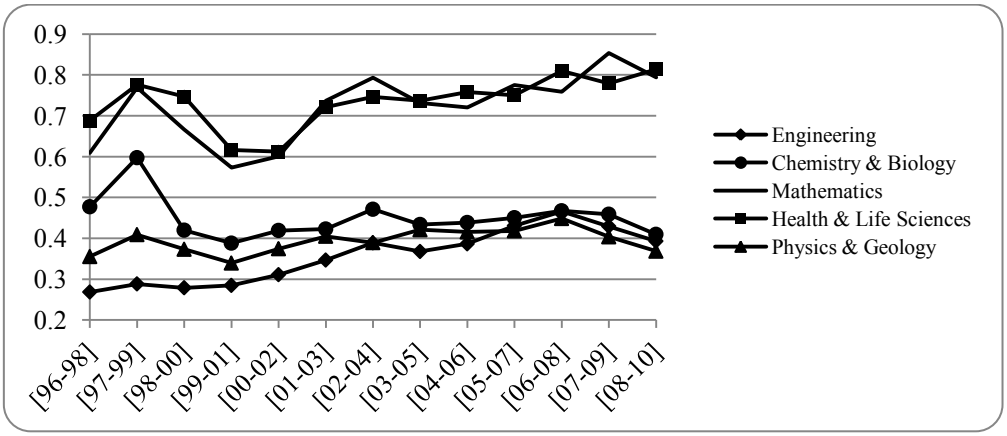


Figure 22. Scientific productivity (Prod3) of the disciplines, [1996-1998] to [2008-2010]

In the final part of this section, the cost of quality and quantity of the funded researchers' papers in different scientific disciplines are compared. For the cost of quality of the papers, we defined "*Quality Indicator*" (QI*)* indicator as follows:

$$QI_{[i,j]} = {avg(funding)_{year=i}}\Big/{avg(impact\ factor)_{year\in[i,j]}}$$

According to the definition of the QI indicator, if QI is higher for a scientific discipline in comparison with the other one(s) it means that articles of the researchers who are working in that scientific field have on average lower impact factor per dollar invested or very high amount of funding, and vice versa. According to Figure 23, the cost of quality is the highest

for the papers of the mathematicians. Except the field of mathematics, the other fields are showing an almost similar trend after the period of [2001-2003] when a considerable drop is observed in all the examined disciplines.

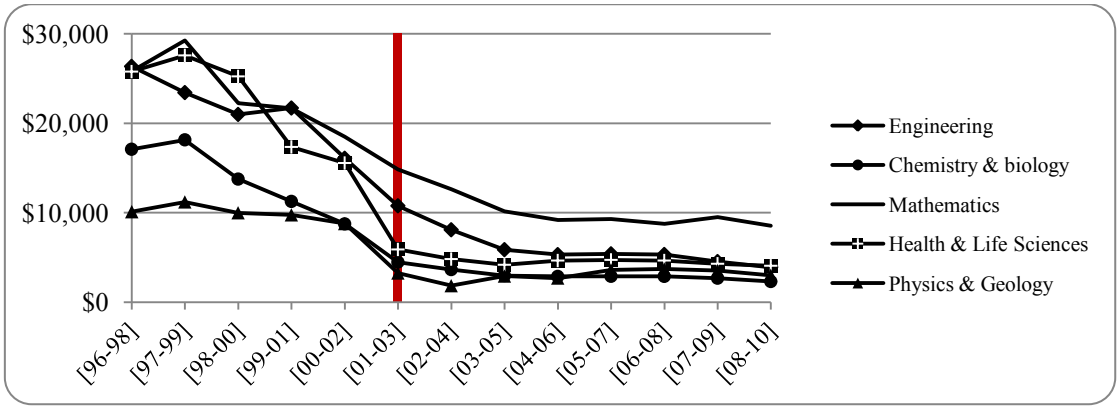


Figure 23. Quality indicator (QI) in different scientific disciplines, [1996-1998] to [2008-2010]

The analysis of the cost of papers (Figure 24) measured by total amount of funding to the number of articles produced reveals that the ratio is the highest for the field of physics and geology. Interestingly, although the cost of works of the engineers was the highest during the beginning periods, a significant drop in the cost is observed after the period of [1999-2001]. Moreover, although after the mentioned period the cost decreased for almost all the examined disciplines, it raises again after the period of [2006-2008]. However, the cost level during the recent years is still lower than its peak level in [1999-2001]. In the next section the impact of different NSERC funding programs on scientific development of the funded researchers is evaluated.

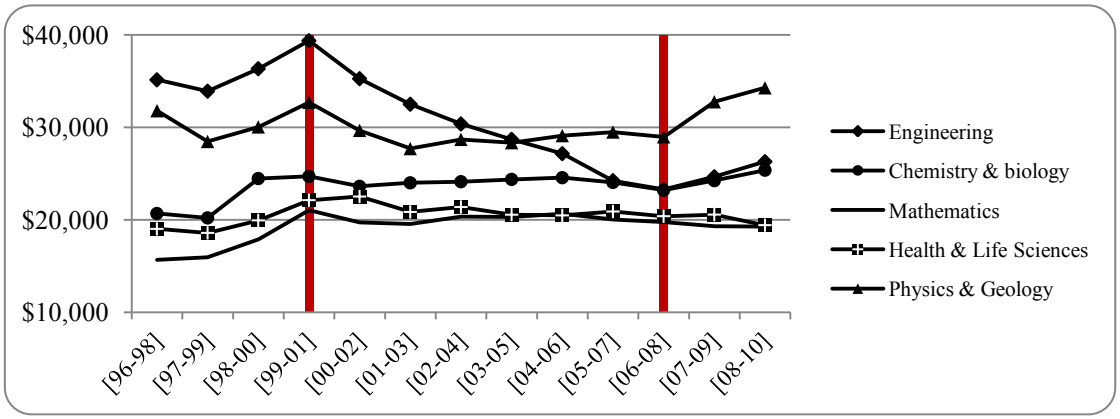


Figure 24. Cost per paper in different scientific disciplines, [1996-1998] to [2008-2010]

5.1.2.3.2 Impact of Different NSERC Funding Programs

135 distinct funding programs were found in our NSERC funding database where the *Discovery Grants Program* was the most frequent one. We selected five frequent funding programs and fetched all the articles that were produced by the researchers funded by the selected programs. Although it could be informative to track different NSERC funding programs, it is difficult to compare the productivity of the researchers funded through different funding programs since most of the researchers have been funded by more than one program. However, analyzing the trends for each of the programs separately could be beneficial in a way that one can understand how effective was the examined program in stimulating scientific activities of the funded researchers.

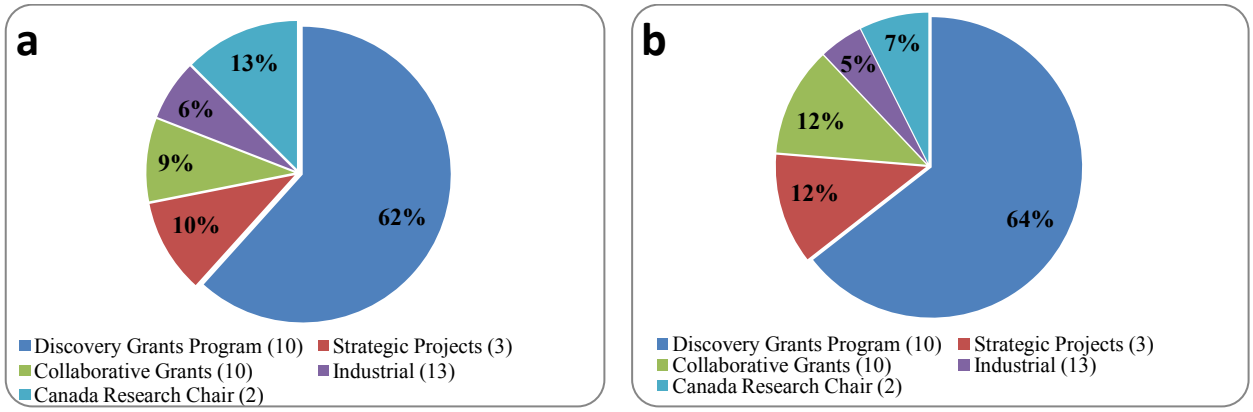


Figure 25. a) Total funding share of the funding programs, 1996-2008, b) Total article share of the funding programs, [1996-1998] to [2008-2010][10]

Each of the five selected main categories of the most frequent NSERC funding programs had some sub-programs, making totally thirty eight funding programs. According to Figure 25-a, discovery grants programs have had the highest share of total funding among the examined programs from 1996 to 2008. Second rank belongs to the Canadian research chair programs that support universities and their affiliated research institutes to improve their research capabilities in order to become world-class research centers. About 10% of NSERC total funding has been dedicated to the strategic projects with an aim to improve the scientific development in selected high-priority areas that will influence Canada's economic and societal position. During the examined period, NSERC has been also supporting researchers' projects through industrial grants, having 6% of the total funding. To foster the

[10] Numbers in parentheses in front of the legends are number of the distinct different subprograms.

scientific collaboration and training activities, 9% has been also dedicated to the funded researchers through the collaborative grants. As it can be observed in Figure 25, article production share of the examined programs (Figure 25-b) is almost corresponding to their share of funding (Figure 25-a), except for Canada research chair program.

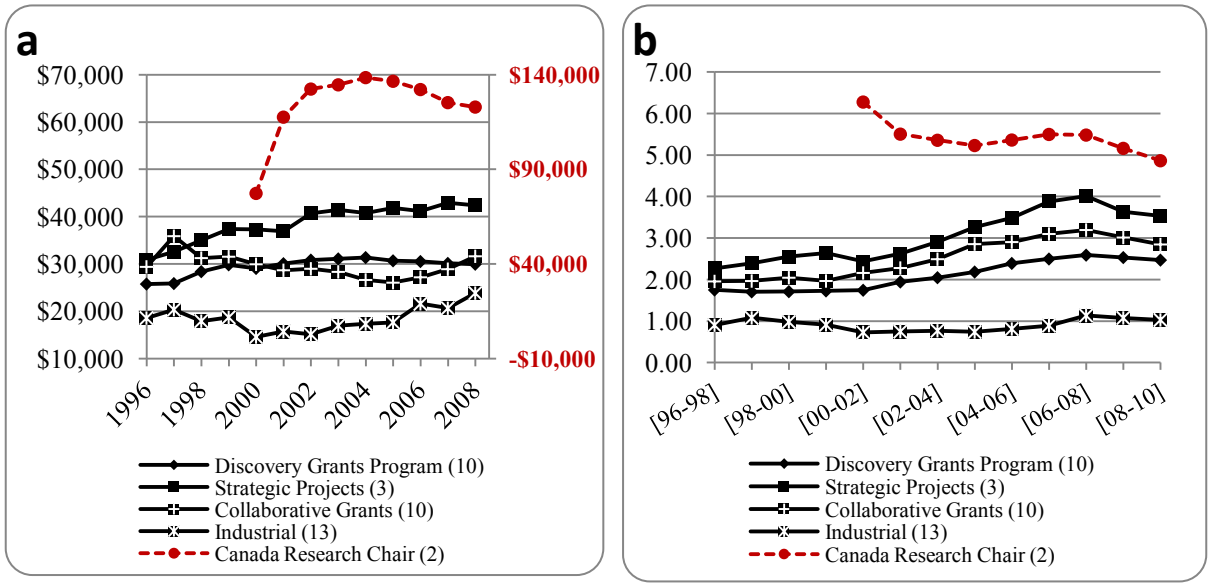


Figure 26. a) Average funding per researcher, 1996-2008, b) Average number of article per researcher, [1996-1998] to [2008-2010]

Analyzing the average amount of NSERC funding per researcher trends in various programs over the examined time interval (Figure 26-a) reveals that the only program that has always followed an increasing trend is the strategic projects funding. It is quite reasonable since according to the definition, this program highly affects the socio-economic macro targets of the country. In addition, although the lowest level of average funding has been allocated through the industrial grants, it has been following an increasing trend after 2000, highlighting the special attention of NSERC to the industrial projects after the mentioned year. However, according to Figure 26-b the increase in the average amount of industrial funding programs has not been resulted in higher productivity of the funded researchers. The average funding allocated to the Canadian research chairs program has been significantly higher than the other examined programs. On the other hand, although the average paper production trend for the researchers funded through the mentioned program is declining, they have produced on average more articles in comparison with the researchers funded through other programs (Figure 26-b). Interestingly, except for the declining curve of

the researchers funded by Canada research chair programs and the almost steady trend of the industrial grants, the average paper production trend for other programs has been almost increasing over the whole period. According to Figure 26-b, productivity of the researchers funded by the discovery grants is the second lowest, ranked after the industrial funding programs. The main reason could be the vast cover of the mentioned program that surely covers less-productive researchers while the other programs targeted more specific areas and/or researchers.

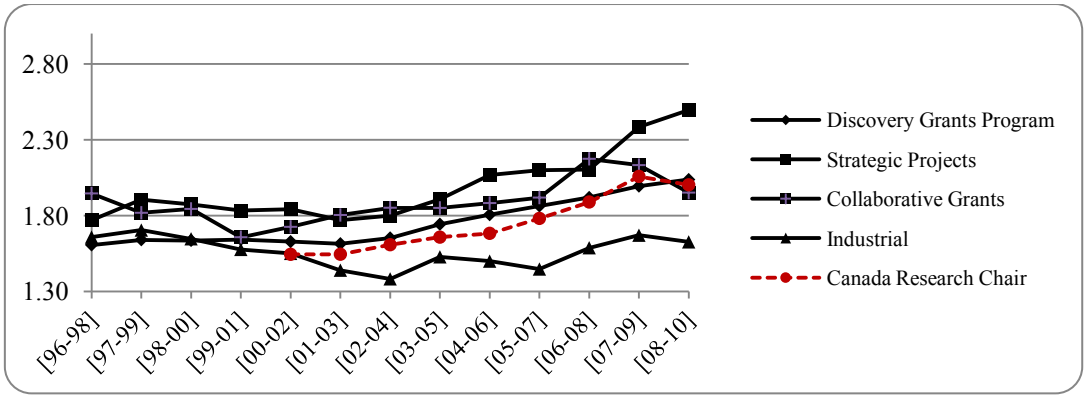


Figure 27. Average journal impact in different funding programs, [1996-1998] to [2008-2010]

Figure 27 depicts average quality of the papers produced by the funded researchers of through various NSERC funding programs measured by the average impact factor of the journals that the papers were published in. As expected, not only the rate of publications for the strategic projects program is high (Figure 26-b), but also the papers are of a high quality following an increasing trend during the examined time interval and be the highest during the last two periods having a considerable difference with the other programs. The quality of the papers for the other programs is almost at the same level except for the ones that have been funded through industrial programs. Moreover, researchers funded through discovery grants programs have been producing relatively low quality papers. This was expected since as it was mentioned a lot of researchers are funded by the aforesaid programs that contain some unproductive or at least less productive researchers. According to the results, researchers funded through Canada research chair program have the lowest quality. Hence,

from Figure 26-a, and 26-b it can be said that the high average investment on this program has been resulted in considerable number of articles, but of low quality. We will further investigate quality of the papers by another indicator that is based on the number of citations received per article.

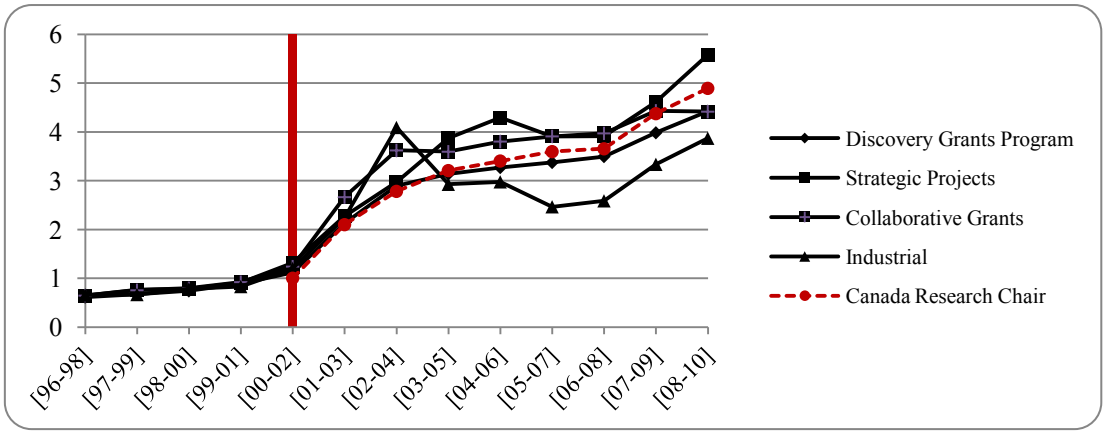


Figure 28. Average number of citations received in various funding programs, [1996-1998] to [2008-2010]

Figure 28 shows the average citations received by papers whose authors were funded through different NSERC funding programs. As it can be seen, in general curves are following an increasing trend except for the industrial funding programs for which a considerable decrease is observed between the periods [2002-2004] to [2005-2007] while it raises again after the period of [2005-2007]. In addition, apart from the industrial funding programs a drastic raise is observed for all the other programs after the period of [2000-2002] that partially confirms the higher rate of average number of citations in the recent years.

Analyzing the multi-authorship trends of the papers produced by the researchers funded by different NSERC funding programs (Figure 29) show that except for the researchers funded by discovery grants programs where the collaborative coefficient is the lowest, the level of multi-authorship for the other funded researchers is almost at the same level following approximately a similar trend. However, in general the trend of multi-authorship is increasing for all the funded researchers especially after 2002. Hence, it can be said that

funded researchers have tried to improve their productivity by expanding their research teams no matter what the source of funding was.

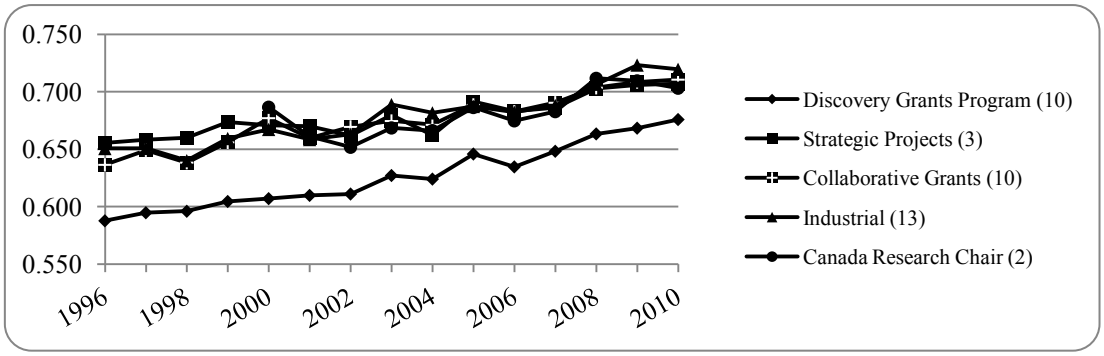


Figure 29. Collaborative Coefficient (CC) in different NSERC funding programs, 1996-2010

In the last part of the analysis we check the efficiency of the funded researchers based on the average amount of funding invested on the quantity (measured by number of articles) and quality (measured by average number of citations) of the articles. Higher QI for a program means that the researchers' articles who were funded through that program have on average lower impact factor per dollar invested or very high amount of funding, and vice versa. As it can be seen in Figure 30, the overall trend of QI is decreasing in all the examined funding programs. In addition, after a considerable drop in the amount of QI from the period of [1996-1998] to [2001-2003], the curves became steady after [2001-2003] except for the Canada research chair program for which the slope is decreasing during the whole examined time interval. In addition, the cost of quality of the papers is the highest for the researchers who were funded through the research chair programs that in this case is mainly due to the high average amount of funding that they have received in comparison with the other funded researchers.

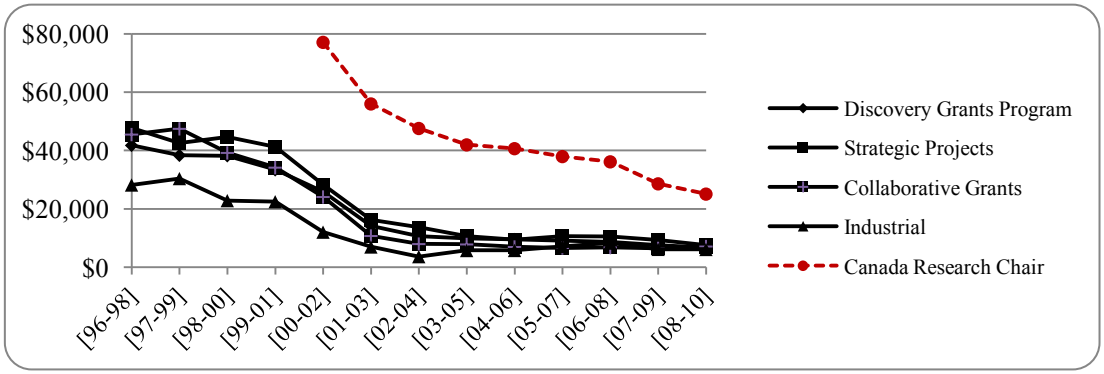


Figure 30. Quality indicator (QI) in different funding programs, [1996-1998] to [2008-2010]

We also checked for the cost of funded researchers' articles. According to Figure 31, not only the cost of articles is the highest for the researchers who were funded through research chair and industrial funding programs but also the trend is almost increasing. This was quite expected since the rate of publication versus the funding allocated was lower for the mentioned programs. Interestingly, the cost of article production for the other programs has been following an almost similar trend during the whole examined time interval. In addition, the curves are following a decreasing trend meaning that on average researchers who were funded through discovery, collaborative, and strategic programs became more efficient in term of average money spent on each article.

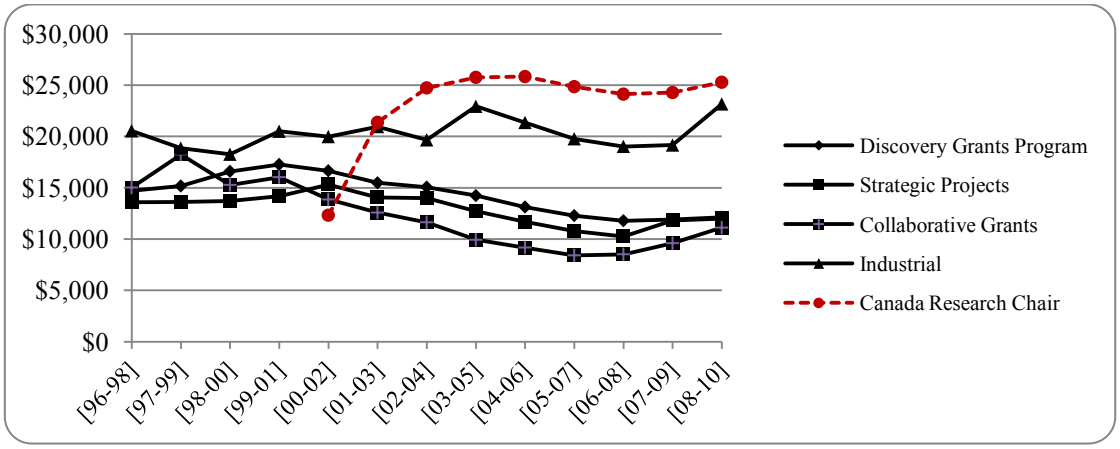


Figure 31. Cost per paper in various funding programs, [1996-1998] to [2008-2010]

5.1.2.4 Conclusion

In this section, scientific development of the NSERC funded researchers from the selected scientific disciplines was investigated and compared. In addition, the impact of some of the most frequent NSERC funding programs was studied. The important role of funding in stimulating scientific activities was partially confirmed. High-priority and more specific funding programs seemed to have resulted in higher productivity of the funded researchers. More specifically, the *strategic projects* funding programs resulted in higher rate of publications and papers of higher quality. In addition, it was the only program where the trend of average funding per researchers was always increasing during the examined period. This may indicate the importance of the defined projects under this type of funding program in improving the societal and economic situation of the country.

A large proportion of the researchers have been funded through the *discovery grants programs*. Although this vast cover secures the input source (even negligible) for the funded researchers, the output of the funded researchers was much lower in comparison with other examined programs, except for the *industrial* programs where the rate of publication was the lowest. One reason could be the inclusion of unproductive or less-productive researchers in this program. The lower quality of the works from the researchers funded by the discovery grants programs is also another proof for this proposition. Regarding the level of multi-authorship, an increasing trend was observed for all the examined programs especially after 2002 meaning that researchers have recently tended more to expand their research teams in an attempt to increase their productivity. In addition, level of the collaborative coefficient was also comparable for all the programs, except for the discovery grants programs.

Analyzing the productivity of the funded researchers in different scientific disciplines revealed that there exist different patterns of co-authorship and productivity. The higher level of multi-authorship in the disciplines of chemistry, physics and health sciences was quite reasonable since some of the projects in the mentioned fields are performed in laboratories where may involve on average more researchers on a project in comparison with engineering, and specially mathematics. In addition, based on the measure introduced in this paper funded researchers from the fields of mathematics and health sciences have been more productive during the examined period. Moreover, an increasing trend was observed for the quality of the papers (measured by the average journal impact factor) in all the disciplines. This raise was more drastic in the field of engineering after the period of [2002-2004]. This high quality may partially justify the lower rate of publication in the field of engineering. On the other hand, the analysis of the citations revealed that a drastic raise was occurred in the average amount of citations received per papers after the period of [2000-2002] that indicates the higher rate of citations in all the disciplines within the recent years.

5.1.2.5 Limitations and Future Work

We were exposed to some limitations in this paper. Firstly, we selected Scopus for gathering information about the NSERC funded researchers' articles. Since Scopus and other similar databases are English biased, non-English articles are underrepresented (Okubo, 1997). Secondly, since Scopus data is less complete before 1996, we had to limit our

analysis to the time interval of 1996 to 2010. Another inevitable limitation related to 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. In addition, in collecting the articles of the funded researchers we assumed that all the funded researchers acknowledge the support of NSERC in the article. This assumption is based on the fact that according to the NSERC guidelines funding has to be acknowledged in the articles of funded researchers. However, it is also probable that some researchers do not acknowledge the source of funding in their papers. Finally, as it was explained before it is hard to assess the net impact of each of the NSERC funding programs since most of the researchers are being funded by more than one program. Hence, assigning the produced paper to the exact source of funding is hard even impossible. For future work various funding councils and sources can be considered in order to compare the scientific productivity of the researchers who were funded through different funding organizations.

5.1.3 Analyzing Scientific Activities of the top Ten Canadian Universities

This section investigates the impact of funding on scientific production of the funded researchers affiliated with the top ten Canadian universities. NSERC funding data in the period of 1996-2010 is considered, and the number of published articles in one-year and three-year time windows is counted as the proxy for the scientific production. In addition, we assess the impact of funding on quality of the funded researchers' papers and their scientific team sizes. Results suggest a positive impact of funding on not only the quantity of the publications but also on the quality of the works and scientific team sizes of the funded researchers.

5.1.3.1 Introduction

Universities and research institutes have an important role in scientific development. In an aim for advancing the scientific position of the country, large amounts of money is being invested annually on research and development (R&D) activities. According to Hagedoorn *et al.* (2000) most of the research projects in the universities are being supported by public funds. Hence, it is essential to define good indicators for evaluating the link between funding and universities' performance, as one of the main drivers of the country's scientific position.

Recent increase in public financing had a considerable impact on the scientific output of the universities (Payne & Siow, 2003; Blume-Kohout, *et al.*, 2009). Arora & Gambardella (1998) analyzed the impact of contractual funding on Italian academic researchers who work in the biotechnology field. They realized the important role of the distribution of funding in a way that more unequal distribution of funds increased the output in the short term. Carayol and Matt (2006) focused on the scientific production of the faculty members of Louis Pasteur University. According to their results, research output is positively influenced by the public contractual funding. In another study, Payne and Siow (2003) analyzed the impact of federal funding on 74 research universities using a panel data set spanning from 1972 to 1998. Their results show a small positive impact of funding on the number of patents while the effect on the number of articles is relatively higher ($1 million leads to 11 more articles and 0.2 more patents). They could not find a significant impact of funding on the quality of the articles measured by number of citations per article. Level of funding allocated to the universities can be considered as an indicator of the quality of the university, even if funding does not improve productivity directly (Blume-Kohout, *et al.*, 2009). In other words, it is expected that the researchers affiliated with higher ranking universities receive higher amount of funding in comparison with other universities.

Scientific activities nowadays know no borders. All researchers worldwide are working together in a global community to improve the level of knowledge. The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006), where financial investment can change the structure of research groups and affect the collaboration among scientists. In two early studies, Beaver and Rosen (1979) and Heffner (1981) found a positive relation between funding and the average number of authors per article. Using questionnaires for gathering data and performing regression analysis, Bozeman and Corley (2004) analyzed the collaboration among a group of scientists affiliated with universities in the U.S. and found a significant positive effect of funding on their collaboration. In another study, Adams *et al.* (2005) confirms the findings of Bozeman and Corley (2004) and found that researchers of top U.S. universities that have larger amounts of federal funding available tend to work in larger scientific groups.

Using a larger and more recent dataset spanning from 1996 to 2010 and by focusing on the top ten Canadian universities, scientific activities of the funded researchers affiliated with the selected universities are analyzed. The rest of the section proceeds as follows: Section 5.1.3.2 describes the data gathering procedure and methodology; Section 5.1.3.3 presents the findings of the analyses; Section 5.1.3.4 concludes while the limitations of the research and some directions for future studies are presented in Section 5.1.3.5.

5.1.3.2 Data and Methodology

The top 10 Canadian universities were selected based on the list on Maclean's website[11] and the scientific performance of their researchers was compared for the period of 1996 to 2010. NSERC was selected as the funding organization to focus on since it is the main federal funding organization in Canada. Almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC (Godin, 2003). The reason for choosing 1996 as the beginning year of the analysis was better coverage of Scopus after 1996. The quality of data before 1996 was lower that might affect the results. Hence, we first collected the funding data from NSERC for the period of 1996 to 2010 that contained information like name of the researcher, his/her affiliation, year, and amount of the award. The extracted data was then refined through employing several automatic cleaning modules coded in JAVA. In addition, team grants were associated to the principal investigator in the original database where we divided the amount equally among the researchers of the team. We hold several interviews with selected researchers in our database where 90% of the interviewees confirm such assumption. The final refined funding dataset contains 75,967 distinct Canadian researchers who received funding from NSERC during the aforementioned period.

As the next step, we searched over Scopus to gather the articles of the NSERC funded researchers for the mentioned period. For this purpose, we searched for all the articles that had acknowledged NSERC funding support within the acknowledgement part of the article. This was a very crucial step in fetching more accurate data that highly influences the findings of the research. The common procedure in similar studies is finding all the articles of the funded researchers that may result in over estimation. Our procedure is based on the

[11] http://oncampus.macleans.ca/education/2013/08/15/23-canadian-universities-make-global-top-500-list/

assumption that according to the NSERC guidelines funding has to be acknowledged in each supported article of the funded researchers. Hence, by our procedure we only count the articles that were produced as the result of NSERC funding, not all the articles of the researcher. This will surely lead to more accurate data and analysis. All the related information such as article co-authors, co-author affiliations, article title, abstract, *etc.* was then extracted. The articles dataset totally contained 130,510 articles and 177,449 authors that acknowledged the NSERC support in the respective article. For counting the articles of the researchers of the examined universities, we considered the affiliation of the author in the year that his/her article has been published.

For evaluating the quality of the papers, SCImago was selected for collecting the impact factor information of the journals in which the articles were published in and the result was integrated into another dataset. SCImago was chosen for three main reasons. Firstly, it provides the journal impact factors for each of the single years of our examined time interval. This enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our articles database. The third reason is about the coverage and quality of SCImago as an open access resource. According to Falagas *et al.* (2008), SCImago covers considerably more journals in comparison with Web of Science which is the base for calculating journal Impact Factor (IF). In addition, SCImago contains a wider variety of countries and languages. Moreover contrary to the journal IF, SCImago's SJR indicator uses different weights to citations depending on the quality of the citing journal. Due to the mentioned advantages, SJR indicator is now considered as a serious alternative to the journal IF.

In this section, we search for relationships between the amounts of funding that NSERC has allocated to the researchers of the top 10 Canadian universities and the scientific productivity of the funded researchers in terms of the number of publications and quality of the papers. We also assess the collaboration among the funded researchers and the impact of NSERC funding on its patterns. We excluded the student grants from the data, since we particularly look for performance of the researchers. Bibliometric analysis is used for the purpose of this study.

5.1.3.3 Results

The top ten Canadian universities was selected based on their rankings in 2013 and the amount of funding allocated to their researchers, their researchers' scientific activities, and the interrelations between funding and scientific performance of the researchers were investigated. Table 6 shows the local and global rankings of the selected Canadian universities in 2013. To assess the impact of funding on the scientific activities of the funded researchers, a three-year (*e.g.* Payne & Siow, 2003) or a five year (*e.g.* Jacob & Lefgren, 2007) time windows have been considered in the literature. We consider one-year and three-year time windows for publications of the funded researchers. As an example of the three-year time window, if the year of funding for a researcher is 1996, we gathered all his/her articles in which NSERC support was acknowledged for the period of 1996 to 1998.

Table 6. Top Ten Canadian universities, 2013

Ranking in Canada	University	World Ranking
1	University of Toronto	28
2	University of British Columbia	40
3	McGill University	58
4	McMaster University	92
5	University of Alberta	101-150
6	University of Montreal	101-150
7	University of Waterloo	151-200
8	Dalhousie University	201-300
9	Laval University	201-300
10	Queen's University	201-300

Source: http://oncampus.macleans.ca/education/2013/08/15/23-canadian-universities-make-global-top-500-list/

According to Figure 32-a, and 32-b, the total funding share and the share of total number of publications for the selected universities does not follow exactly the order of their rankings. However, the top three highest ranking universities (*i.e.* University of Toronto, University of British Columbia, and McGill University) plus University of Alberta (ranked 5th) and Waterloo University (ranked 7th) are the top 5 universities considering both their total share of funding and their total share of publications. In addition except for Laval

University, it seems that there is a positive relation between the amount of funding allocated to the universities and the number of articles that they have produced.

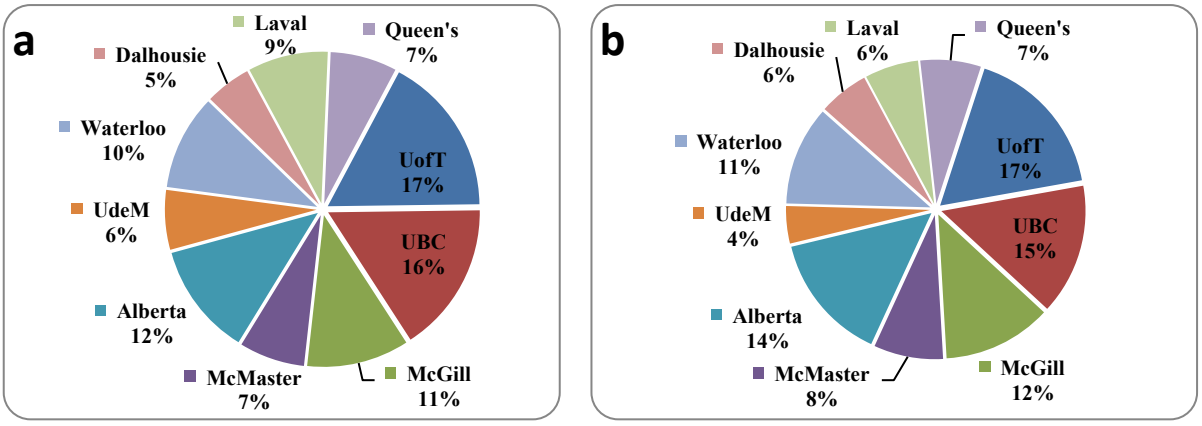


Figure 32. a) Total funding share of the top 10 Canadian universities, 1996-2010, b) Total number of articles share of the top 10 Canadian universities, 1996-2010

To have a better picture of the relation, we take the number of researchers into the account. As it can be seen in Figure 33-a, the average share of funding for the examined universities is almost at the same level (ranging from 9% to 11%). This may partially reflect the special and fair attention of NSERC to the researchers affiliated with the top ten Canadian universities. Despite having quite comparable share of funding, the share of publication for the examined universities differs more, ranging from 6% for UdeM to 12% for Alberta University (Figure 33-b). However, it seems that the language factor might have played a minor role here since the two universities that have the lowest share of publications are French speaking universities, namely UdeM and Laval University. In other words, the researchers of the two mentioned universities may have also some publications in French that are not being counted in our analysis since we use Scopus as the source for researchers' articles.

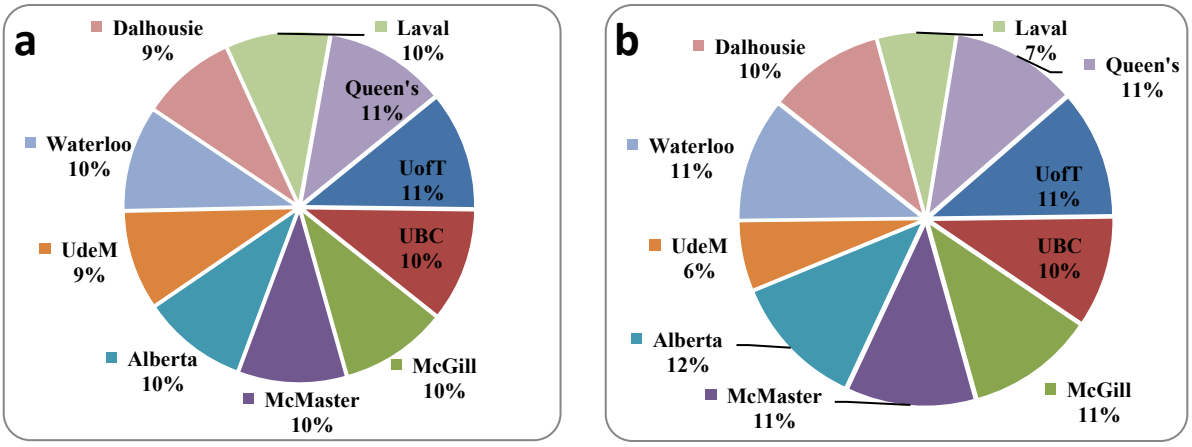


Figure 33. a) Total average funding share per researcher in the top 10 Canadian universities, b) Total average share of article production per researcher in the top 10 Canadian universities

Apart from the total share of funding and number of articles, analyzing their trends during the examined time interval can be also informative. According to Figure 34-a, the average funding per researcher for the top ten Canadian universities has been always increasing without any steady period (the only exception is Queen's University that will be discussed later). Comparing this finding with the overall funding trends, it can be said that although there exists some steady (or with little increase) NSERC funding periods for Canadian provinces, the funding allocated to the top universities has not been decreased. Hence during the low budget periods, it seems that NSERC decreased the funding of the less productive research institutes and universities and tried to constantly increase the budget of the high ranking universities in an attempt to boost the scientific development. For Queen's university a drastic jump in average grants is observed after 2007. According to the NSERC funding database a considerable amount of funding has been allocated to the researchers of Queen's University from 2007 to 2010 through NSERC's "*cooperative activities*" program, ranging from around $9.2 Millions in 2007 to more than $6.6 Millions in 2010. Later, we will further investigate the impact of this special support of cooperative activities on the formation and collaboration patterns of the Queen's University researchers. Excluding the curve of Queen's University, researchers of University of Toronto have been receiving the highest amount of average funding during the examined period while the lowest average funding has been allocated to the researchers of Dalhousie University.

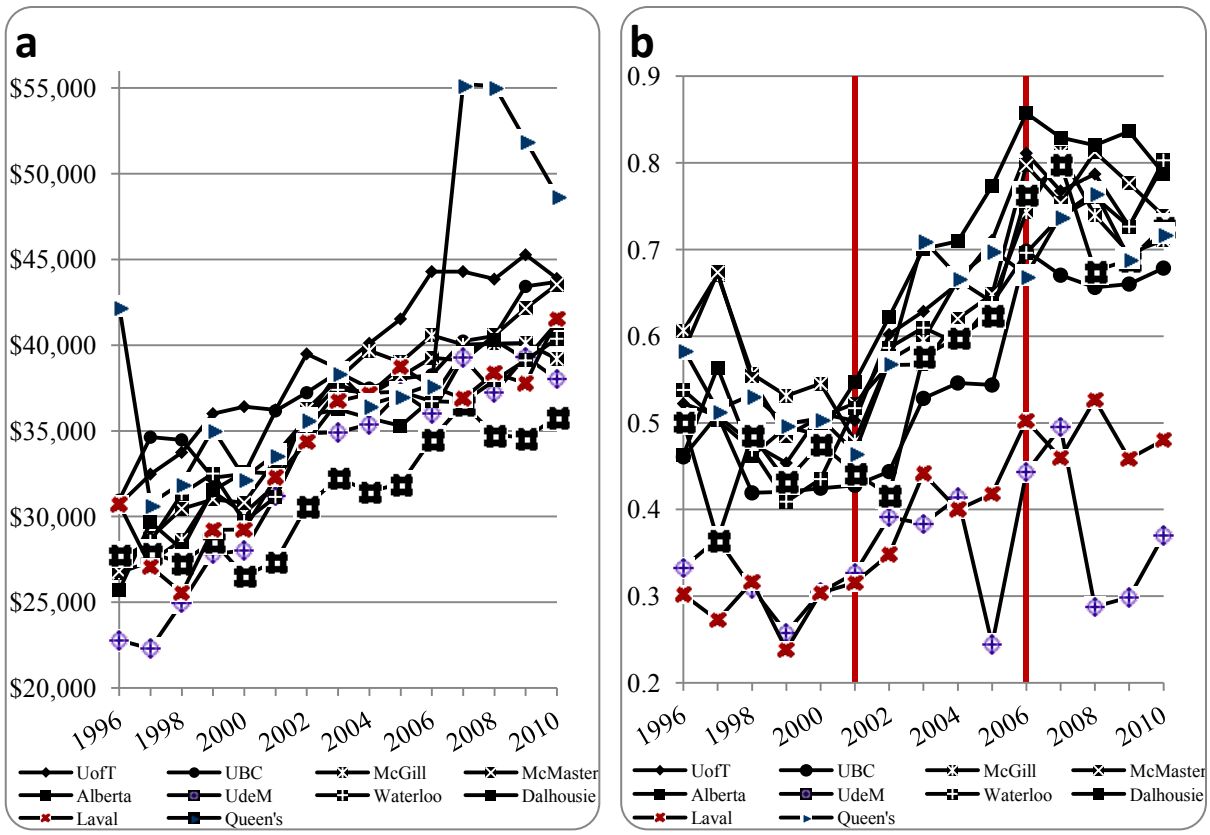


Figure 34. a) Average funding per researcher in the top 10 Canadian universities, 1996-2010, b) Average number of articles per researcher in the top 10 Canadian universities, 1996-2010

Figure 34-b depicts the trend of average number of articles per researcher in the selected universities. The trend of researchers' average number of publications can be divided into three different periods, highlighted by the vertical red lines in the respective figure. From 1996 to 2001 and from 2006 to 2010 we see a slightly decreasing overall trend (steady in some cases). However, from 2001 till 2006 a significant increase is observed in the number of articles per researcher. Comparing this finding with Figure 34-a, interestingly we see that the curve of average funding is steeper during the mentioned period. Hence, it seems that higher level of funding available has positively influenced the average productivity of the researchers. In addition, in line with our findings from Figure 33 the lowest rate of average article production belongs to the researchers of UdeM and Laval universities where the gap between the mentioned universities and the other ones becomes bigger as we move forward toward the time axis. University of Alberta is producing the highest number of publications while the average amount of funding allocated to its researchers is not even among the top four during the whole examined time interval.

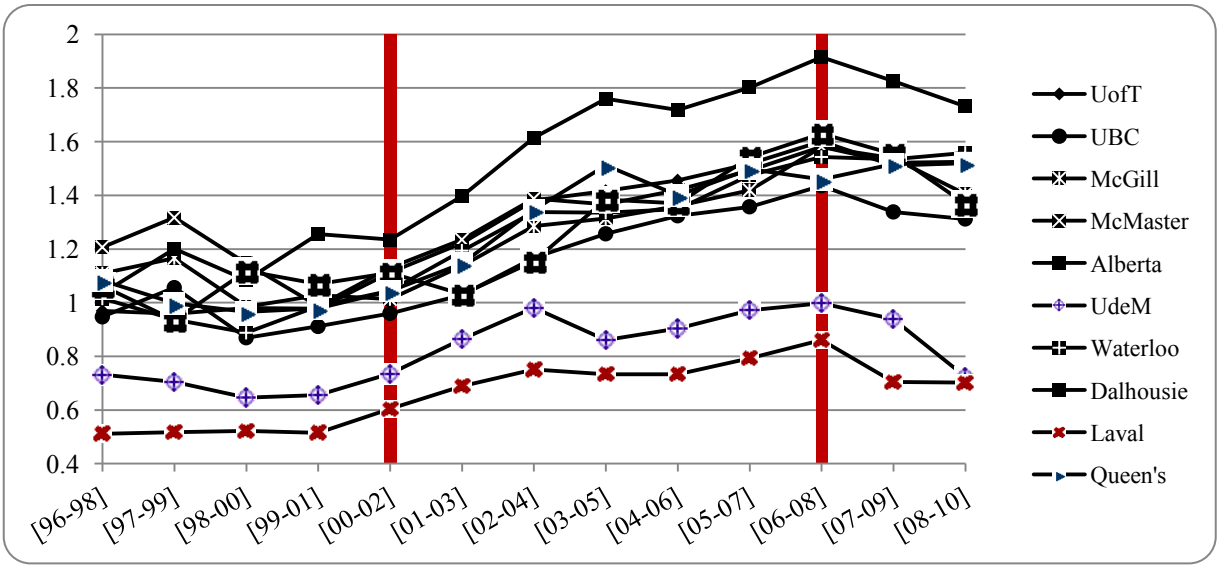


Figure 35. Average number of articles per researcher in the top 10 Canadian universities, [1996-1998] to [2008-2010]

To see the impact of funding on the number of articles more accurately, we also considered a 3-year time window for the publications. According to Figure 35, three regions same as Figure 34-b with the same explanations can be observed for the examined universities. However, one difference is the first region that ends in the period of [2000-2002] in the case of the three-year time window. In addition, the drop in the productivity of the funded researchers during the last period (from [2006-2008] to [2008-2010]) is more sensible in Figure 35. Hence, one reason for this drop could be the almost steady trend (slightly increasing for some universities) of the average funding after 2006. We will further investigate this issue by considering other factors (*e.g.* quality of the papers, collaboration patterns among the scientists, *etc.*).

Apart from the number of publications, one should also consider the quality of the works that have been produced. This may help to have a better picture of productivity and efficiency of the funded researchers. Figure 36-a depicts the trend of average impact factor of the journals in which the articles were published in the same year that the funded researcher has received funding. Figure 36-b shows the same indicator calculated for the articles published in a three-year time window, beginning with the year that the researcher has received grants. As an example, we collected all the articles within the period of 1996 to 1998 for the researchers who were funded in 1996 and calculated the average impact factor of the journals that the articles were published in.

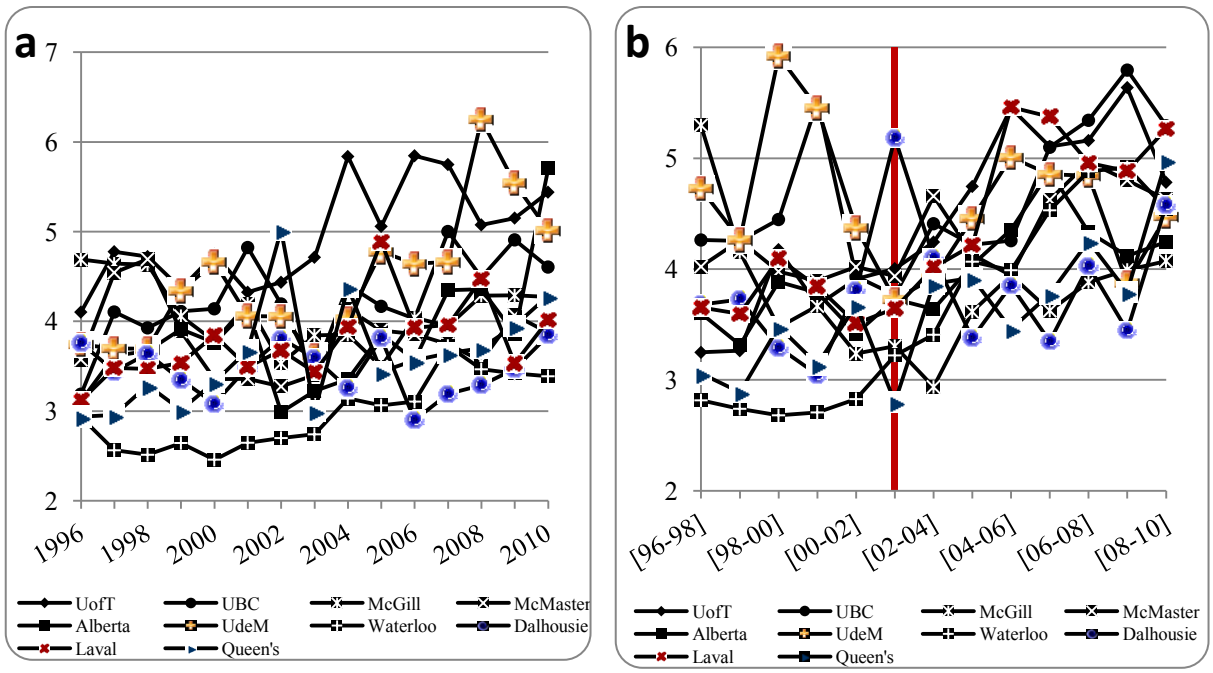


Figure 36. a) Average impact factor of articles in the top 10 Canadian universities in the same funding year, 1996-2010, b) Average impact factor of articles (3-year time window) in the top 10 Canadian universities, [1996-1998] to [2008-2010]

According to Figures 36-a, and 36-b, it can be said that the trend of quality of the papers has followed an increasing trend recently. This observation is clearer in Figure 36-b where after the period of [2001-2003] we see steeper curves for most of the examined universities. According to Figure 36-a, UdeM, UofT, and UBC have been the top three universities during the past five years. Considering the three-year publications, just Laval University takes the place of UdeM among the top three universities. This is quite interesting since it seems that French Speaking universities (UdeM, and Laval University) that showed low average rate of publications (Figure 34-b, and Figure 35) preferred to publish in higher quality journals. In addition, comparing Figures 35, and 36-b an almost steady (decreasing in some years) trend is also observed for the quality of the papers before [2001-2003]. Hence, it seems that the steady trend of funding has influenced both quantity and quality of the papers before the mentioned period. However, the drop in the number of publications that can be observed after [2006-2008] in Figure 35 is not so obvious in Figure 36-b. Hence, although the trend of average funding became slighter after 2006, it seems that some researchers tended to decrease the number of publications but still preferred to continue publishing in relatively high quality journals. However, from the fluctuations in the curves after the period of [2006-2008] in Figure 36-b it could be possible that if funding continues to follow a not

increasing trend it may finally influence the quality of the works of the researchers negatively.

We also investigated the trend of average citation counts in a three-year time window for the selected universities. The results are shown in Figure 37. Although a steady trend (slightly increasing in a few cases) is seen for almost all the examined universities before the period of [2000-2002], interestingly, a drastic increase is observed in the average number of citations after the mentioned period which is almost in line with our findings from Figure 36-b. This partially empowers the validation of the pervious discussion made on Figure 36.

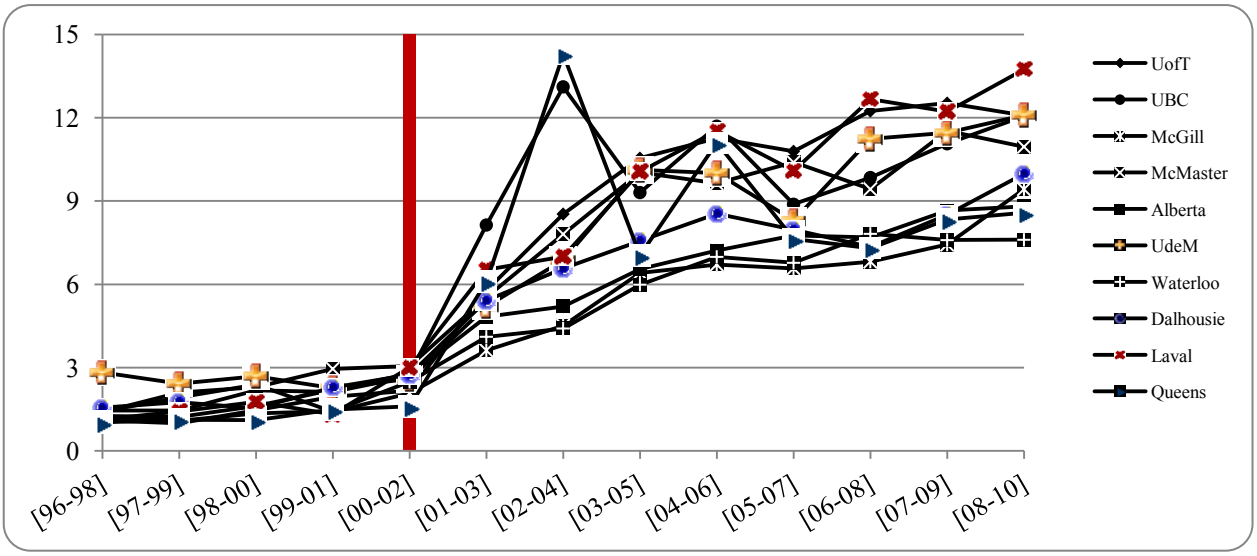


Figure 37. Average number of citations (3-year time window) in the top 10 Canadian universities, [1996-1998] to [2008-2010]

In the rest of the section, the impact of funding on scientific collaboration patterns of the funded researchers is evaluated. Higher level of funding can enable researchers to expand their scientific activities that may result in higher scientific production. To analyze the collaboration patterns of the funded researchers, we used average number of authors per paper (APP) as the proxy of the team size. Figure 38 shows the trend of APP for the publications of the funded researchers in a three-year time window. As it can be seen the slope of the curves becomes slightly higher after the period of [2001-2003]. Interestingly, this point is where we see an increase in the average number of publications (Figure 34-b). Moreover, from 2001 to 2006 we also see a drastic increase in the average amount of

funding, which decreased a bit after 2006 till 2010. Hence, it can be said that higher level of funding might have enabled researchers to expand their scientific activities (*e.g.* through getting involved in larger or new projects, collaborating with new partners, *etc.*) that caused an increase in the average number of publications. Interestingly, after [2001-2003] we also see an increase in the average impact factor of the journals (Figure 36-b). Therefore, it seems that there is a chain relation among funding, quantity and quality of the articles, and average team size of the funded researchers.

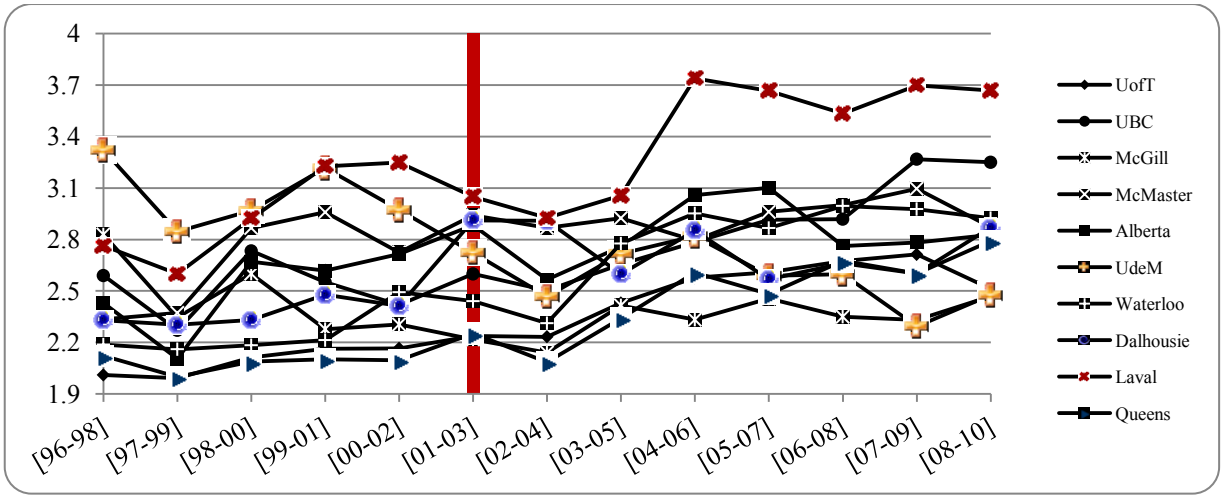


Figure 38. Authors per paper (APP) (3-year time window) in the top 10 Canadian universities, [1996-1998] to [2008-2010]

In addition, it is interesting that, in general, researchers from all the selected universities (except UdeM) tried to be gradually more involved in larger scientific teams as we move forward toward the time axis, especially after the period of [2001-2003]. This can partially confirms the importance of the well-formed and reasonably large scientific teams among the researchers as the nature of the science is getting more complex, modern, and inter-disciplinary day by day.

5.1.3.4 Conclusion

In this section, we particularly focused on the top ten highest ranking Canadian universities and analyzed the performance and collaboration patterns of their researchers in respect to the amount of funding that they have been receiving from NSERC during the time interval of 1996 to 2010. According to our results, the average share of funding per researcher for all the examined universities were almost at the same level (9% to 11%). In

addition, except the French speaking universities (UdeM and Laval universities with the share of 6% and 7% respectively) the average share of publications per researcher for the rest of the universities was also at the same level (10% to 12%). As discussed, the language factor could also play a minor role here in a way that researchers from Quebec may also publish in French language that is not counted in Scopus. Except the mentioned two universities, the share of funding and publications were quite comparable for the other eight universities. Interestingly, the NSERC average funding curves for the examined universities follows an increasing trend almost during the whole time interval. In other words, NSERC has constantly increased the average funding allocated to the high ranking universities may be due to their higher scientific performance in comparison with the other Canadian universities.

Our results suggest that there is a positive relation between level of funding and the productivity of the funded researchers affiliated with the top ten Canadian universities. Considerable increase in the average amount of funding for the period of 2000 to 2006 is concurrent with the boost in the number of publications during the same period. Moreover, whenever we see a declining or steadier trend of the average amount of funding (especially after 2006), the average number of publications also decreases. This was expected since funding has been acknowledged as one of the most determinant factors in stimulating scientific activities (Martin, 2003). Having more money available can enable researchers to get access to better research resources (Lee & Bozeman, 2005), that might be resulted in higher productivity. Our finding is in line with several studies that focused on the universities performance and found a positive relation between funding and scientific performance (*e.g.* Payne & Siow, 2003; Carayol & Matt, 2006; Blume-Kohout, *et al.*, 2009).

Moreover, the impact of funding on quality of the papers of the funded researchers was analyzed. It was interesting to observe a positive impact of funding on quality of the papers, measured by the average impact factor of the journals that articles were published in. Specifically, it seems that the raise in the amount of average funding after 2000 was one of the most significant reasons of the increase in the quality of the papers after the period of [2001-2003]. One reason of such relation could be the decline in the quality indicator after the period of [2006-2008] which is also concurrent with the decline of funding. More

interestingly a raise was also observed for the team size of the funded researchers of the top ten Canadian high ranking universities exactly after the period of [2001-2003]. Hence, it can be said that higher level of funding allocated to the researchers enabled them to expand their team sizes through cooperating with new partners and most probably experts of the field that has resulted in both higher quantity and quality of the papers. Two similar studies, *i.e.* Godin (2003) and Payne and Siow (2003), found no impact of funding on quality of the papers of the funded researchers. Godin (2003) focused on NSERC funding as the input source and Payne and Siow (2003) analyzed the performance of 74 research universities against deferral funding.

Although the increase in the funding level has been followed by higher productivity and larger team sizes in most of the periods of the examined time interval, one should notice that there may exists other factors rather than funding (*e.g.* research policies and priorities, cultural issues, *etc.*) that could have influenced the scientific development. Hence, complementary analysis is needed in this regard to make any final conclusions. However, it seems that funding, directly or indirectly, influences the scientific activities of the researchers affiliated with the top ten Canadian universities in a way that higher funding has encouraged researchers to produce more articles through getting involved in larger scientific teams while paying more attention to the quality of their work.

5.1.3.5 Limitations and Future Work

We were exposed to some limitations. Firstly, Scopus that was selected for gathering information about the NSERC funded researchers' articles is English biased, non-English articles are underrepresented (Okubo, 1997). Secondly, since Scopus data is less complete before 1996, we had to limit our analysis to the time interval of 1996 to 2010. Another inevitable limitation related to 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. In addition, in collecting the articles of the funded researchers we assumed that all the funded researchers acknowledge support of NSERC in the article. This assumption is based on the fact that according to the NSERC guidelines funding has to be acknowledged in the articles of funded researchers. However, it is also probable that some researchers do not

acknowledge the source of funding in their papers. Moreover, some journals might ask the author(s) to remove the acknowledgement from the article. We could not also count these articles in our analysis.

Different scientific disciplines follow different patterns in publishing articles, collaborating with other researchers, or even getting and allocating grants to the tasks. Hence to better examine scientific productivity and efficiency, a future work direction could be assessing the impact of funding on the rate of publications for different scientific disciplines of the universities separately. In addition, the impact could be separately analyzed for different types and programs of funding, and also other funding councils can be considered as the source of funding data. This kind of analyses and comparing the efficiency of different funding organizations may help the decision makers to draw a better picture of the performance of the researchers affiliated with the high ranking universities.

5.2 Statistical Analysis

Four papers were produced using statistical and social network analyses that are presented in separately in this section. The first paper, titled "*On the Impact of Funding on Scientific Production: A Statistical Analysis Approach*", evaluates the impact of a number of influencing factors on scientific productivity of the funded researchers at the individual level. Section 5.2.1 is dedicated to this paper. Section 5.2.2 discusses the results of the second paper, titled "*On the Impact of the Small World Structure on Scientific Activities*". This paper tests the existence of the small world property in the collaboration network on the NSERC funded researchers and evaluates its impact on productivity, quality of the works, and scientific team size of the researchers. Section 5.1.3 belongs to the third paper, titled "*How the Influencing Factors Affect Researchers' Collaborative Behavior?*". This paper focuses on the collaboration network of the funded researchers and employs time related statistical models to estimate the impact of the influencing factors on the network structure variables. The title of the last paper is "*How to Get more Funding for Research*" that is presented in section 5.1.4 and discusses the impact of several important factors on the amount of funding that is allocated to the researchers.

5.2.1 On the Impact of Funding on Scientific Production: A Statistical Analysis Approach

The impact of funding on the scientific production of the funded researchers is statistically investigated in this section. Number of published articles is counted as the proxy for the scientific production and the average number of citations is considered as the measure for quality of the papers. Time related statistical models for the period of 1996 to 2010 are estimated to assess the impact of funding and other influencing factors on the quantity and quality of the scientific output of individual funded researchers. Results confirm a positive impact of funding on the quantity and quality of the publications.

5.2.1.1 Introduction

Billions of dollars are being annually spent on the research and development (R&D) activities through the federal funding agencies. Universities, colleges, and research institutes are the key players in knowledge production (Gulbrandsen & Smeby, 2005). In 2013, the Natural Sciences and Engineering Research Council of Canada (NSERC), as one of the major Canadian federal granting agencies, invested more than one billion dollars on research by funding more than 29,000 students, over 11,000 university professors, and about 2,400 Canadian-based companies (NSERC, 2013). Other federal funding agencies of Canada (*e.g.* Canadian Institute of Health Research (CIHR)[12], and Social Sciences and Humanities Research Council of Canada (SSHRC)[13] are also supporting the researchers in an aim to improve the socio-economic situation of the country.

Research requires appropriate amount of investment enabling the researchers to purchase the required equipments, tools, or to be able to cooperate with other experts in the field. Hence, research is often expensive (Gulbrandsen & Smeby, 2005). On the other hand, better access to the funding resources can make prominent researchers more productive bringing gradually more credit and disproportionate resource to them. This process is called "*credibility cycle*" in the literature (Latour & Woolgar, 1979). Hence, wise funding allocation process and well-established researchers' performance evaluation system are required.

[12] For more information see: http://www.cihr-irsc.gc.ca/e/193.html
[13] For more information see: http://www.sshrc-crsh.gc.ca/home-accueil-eng.aspx

Evaluating the relation between research input (*e.g.* research funding) and the quantity (*e.g.* number of publications) and quality (*e.g.* number of citations) of the research output has been a challenging issue for policy makers. A number of techniques (*e.g.* bibliometrics, statistical analysis) have been used for this purpose (King, 1987). In an early case study performed by McAllister and Narin (1983) for the National Institute of Health (NIH) the relation between NIH's funding and number of publications of the U.S. medical schools was investigated. Using bibliometric indicators they found a quite strong relationship between funding and the number of papers published. A few other studies investigated the effect of funding on the output of medical (health) schools (programs) (*e.g.* Lewison & Dawson, 1998; Jacob & Lefgren, 2007; Albrecht, 2009).

Analyzing the impact of financial investment on scientific production at cross-country level has also attracted scientists' attention (*e.g.* Leydesdorff & Wagner, 2009; Crespi & Geuna, 2008). Shapira and Wang (2010) investigated the impact of nanotechnology funding. They used Thomson Reuter's database for the period of August 2008 to July 2009 and used very basic bibliometric indicators to give a general picture of countries which are working in nanotechnology field. They argued that as an impact of large investment that has been made, China is getting closer to the U.S. in terms of the number of publications but Chinese papers still have lower quality in comparison with the Americans and Europeans.

Several studies investigated the impact of funding on the performance of academic researchers. Payne and Siow (2003) analyzed the impact of federal funding on 74 research universities. Employing a regression analysis on a panel data set spanning from 1972 to 1998, they investigated the effects of funding on the articles published and patents issued by the researchers. Their results show a small positive impact of funding on the number of patents while the effect on the number of articles is relatively higher. In an econometric evaluation of the impact of funding composition on agricultural productivity, Huffman and Evenson (2005) used annual data for 48 U.S. states from 1970 to 1999. They found a significant negative impact of the federal competitive grant funding on the productivity of public agricultural researchers. A number of other studies have also studied the effect of funding on the performance of academics (*e.g.* Gulbrandsen & Smeby, 2005; Beaudry & Allaoui, 2012).

The evaluation of research performance in Canada started attracting the attention of the policy makers recently. In Canada, scientific articles have been recognized as the main output of researchers and universities (Godin, 2003) and bibliometrics has been mostly used for scientific evaluation purposes. Gingras (1996) in a report to the Program Evaluation Committee of NSERC discussed the feasibility of bibliometric evaluation of the funded research. Following his study, a few other Canadian researchers used bibliometrics for analyzing the funding impact (*e.g.* Godin, 2003; Campbell, *et al.*, 2010; Campbell & Bertrand, 2009) that mostly found a positive relation between funding and productivity. However, the datasets that were used for the analysis were limited in most of the cases and simple indicators were used for the analysis. In addition, the analyses were not done at the individual level of the researchers. These gaps call for a more comprehensive study in Canada.

Although most of the studies in the literature have found a positive relation between funding and the rate of the publications regardless of intensity of the relation (*e.g.* Godin, 2003; Payne & Siow, 2003; Jacob & Lefgren, 2007), there also exist some studies that found no significant relation (*e.g.* Beaudry & Allaoui, 2012[14]; Carayol & Matt, 2006) or even a negative impact (*e.g.* Huffman & Evenson, 2005). Hence, the results are inconsistent and the relation needs further investigation. This research uses a larger and more recent data spanning from 1996 to 2010 and applies a unique procedure for collecting the funded researchers' articles more accurately. It employs several statistical models and new to the field independent variables to comprehensively study the impact of the influencing factors on scientific production in Canada. The remainder of the section proceeds as follows: Section 5.2.1.2 presents the data, methodology and the general models; Section 5.2.1.3 presents the empirical results and interpretations; Section 5.2.1.4 concludes; and Section 5.2.1.5 discusses the limitations.

[14] They found no impact of private funding but positive impact of public funding.

5.2.1.2 Data and Methodology

5.2.1.2.1 Data

Three data sets of funding, funded researchers' publications, and articles' quality were integrated in this research. NSERC was selected as the focal funding organization of this research. The main reasons for choosing NSERC was its role as the main federal funding organization in Canada, and the fact that almost all the Canadian researchers in natural sciences and engineering receive at least a basic research grant from NSERC (Godin, 2003). As the first stage, NSERC funding data were extracted for the period of 1996 to 2010. Elsevier's Scopus was selected as the source of scientific publications. It provides the necessary data on the articles, *e.g.* co-authors, their affiliations, year of publication, *etc*. As the second step, all the scientific articles that had acknowledged the support of NSERC in the body of the paper were extracted for the period of 1996 to 2010. 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. This must have resulted in an over-estimation of the number of articles, as researchers usually use several sources of funding. The acknowledgement-based search was based on the assumption that all the NSERC grantees acknowledge the source of funding in the article. We validated this assumption through holding interviews with 30 randomly selected researchers in our database. The reason for selecting the time interval from 1996 to 2010 was low data quality of Scopus before 1996. In total, 120,439 articles authored by 36,124 distinct authors from 1996 to 2010 were collected. SCImago was selected for collecting the impact factor information of the journals in which the articles were published in. We used this data as a proxy of the quality of the papers. SCImago does not provide the impact factor data before 1999 hence we considered 1999 data for the articles published in the period of 1996 to 1999. For the rest of the articles we used the impact factor of the journal in the year that the article was published in. Through an automatic careful examination of the first names, surnames, initials, and affiliations, Scopus ID of the funded researchers was then extracted that was used to integrate the mentioned data sets.

5.2.1.2.2 Model Specification and Variables

This research investigates the impact of some of the influencing variables on the quantity and quality of the publications of the NSERC funded researchers. The models and variables that were used for each of the estimations are presented in the following sections. STATA 12[15] data analysis and statistical software wass used to estimate the models.

5.2.1.2.2.1 Quantity of the Publications Model

Since the purpose of this research is to study the impact of funding and past productivity related variables on the scientific productivity of the funded researchers we consider the number of articles in a given year as the dependent variable (*noArt*). Our dependent variable is therefore a count measure. Hausman *et al.* (1984) proposed the Poisson model for a count measure. Although the best matching regression model is Poisson, in reality it is rare to satisfy the Poisson assumption on the actual distribution of a natural phenomenon, because most of the time an over-dispersion or under-dispersion is detected in the sample data. This causes the Poisson model to underestimate or overestimate the standard errors and thus results in misleading estimates for the statistical significance of variables (Coleman & Lazarsfeld, 1981). According to Hausman *et al.* (1984), in order to obtain robust standard errors correcting the estimates binomial regression can be employed. Therefore, we used negative binomial regression to estimate the number of papers published in a given year by an individual. The regression model in the reduced form is as follows:

$$noArt_i = f(avgFund3_{i-1} + avgIf3_{i-1} + noArt_{i-1} + avgCit3_{i-1} + avgTeamSize_i + careerAgei + dProvincei + dInsti + dFundProgi \quad (1)$$

In the model, $avgFund3_{i-1}$ is the average amount of funding that the researcher has received over the past three years. In the literature three-year (*e.g.* Payne & Siow, 2003) or five year (*e.g.* Jacob & Lefgren, 2007) time windows have been considered for the funding to take effect. We considered both for our model and found that the three-year time window is better suited. We calculated the average impact factor of the journals that the author has published articles in ($avgIf3_{i-1}$) for a three-year time interval as a proxy for the quality of his/her papers. As another measure for the quality of the papers, we also added $avgCit3_{i-1}$

[15] For more information see: http://www.stata.com/stata12/

variable to the model that is the average citations for the articles in a three year time window. $AvgTeamSize_i$ represents the average number of co-authors in an author's papers in a given year. We also considered the past productivity of the funded researcher represented by $noArt_{i-1}$ in the model.

In general, older researchers can be more productive (Merton, 1973; Kyvik & Olsen, 2008) due to several factors *e.g.* better access to the funding and expertise sources, more established network, better access to modern equipments, *etc*. 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. We also added dummy variables to the models: $dProvince_i$ represents the Canadian provinces, $dInst_i$ for the fact that whether the funded researcher is affiliated with academia or non-academia environments, and $dFundProg_i$ for representing various NSERC funding programs.

5.2.1.2.2.2 Quality of the Publications Model

To investigate the impact on the quality of funded researchers' papers, we considered the average amount of citations for all the articles of a funded researcher in year *i* as the dependent variable ($avgCit_i$). The following regression model (reduced form) is used:

$$avgCit_i = f(avgFund3_{i-1} + avgArt3_{i-1} + avgIf3_{i-1} + avgCit3_{i-1} + avgTeamSize_i + careerAge_i + dProvince_i + dInst_i + dFundProg_i) \quad (2)$$

The definition of the variables are the same as the ones for model (1) except for $avgArt3_{i-1}$ that is the average number of publications for a funded researcher in the period of [*i-1,i-3*] if the research has been funded in year *i*. This variable indicates the past productivity of the research in terms of his/her average number of publications in a three-year time window. In the following section results are presented and discussed.

5.2.1.3 Results

Results of the analyses are presented in two sections. In the first section, the results of the visualization analysis and descriptive statistics are stated. RapidMiner[16] software was used for the visualizations. The second section discusses the results of the regression analysis.

5.2.1.3.1 Visualization and Descriptive Analysis

Data visualizations are used to find some preliminary patterns in the data. Figure 39 shows the trend of funding over the examined period. We adjusted the amount of total funding based on the constant Canadian dollar in 2003 to remove the general effects of expenditure increase. As it can be seen, a significant raise is observed form 2001 to 2007. After 2007, the trend of inflation adjusted total funding is almost constant maintaining its level around $900 million.

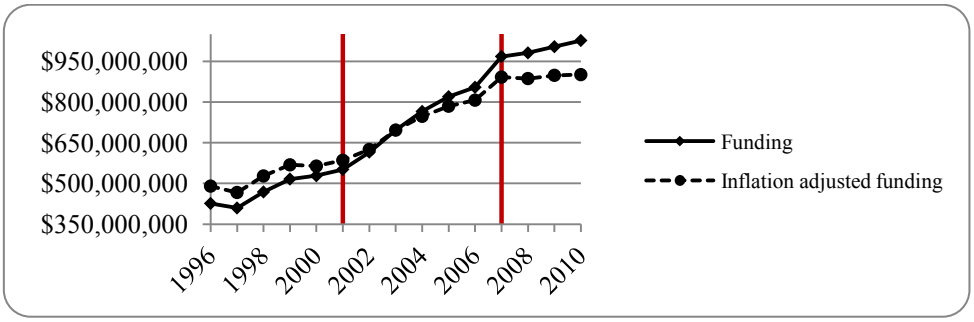


Figure 39. Trend of total funding and inflation adjusted funding, 1996-2010

In Figure 40 the trend of average inflation adjusted funding invested on each article produced is depicted. According to the figure, it can be seen that the cost of articles has been on average decreasing after 1999. In other words, with the same level of funding available funded researchers have produced on average more articles. Hence, this can be a partial indicator of the raise in the number of publications especially after 1999.

[16] For more information see: http://rapidminer.com/

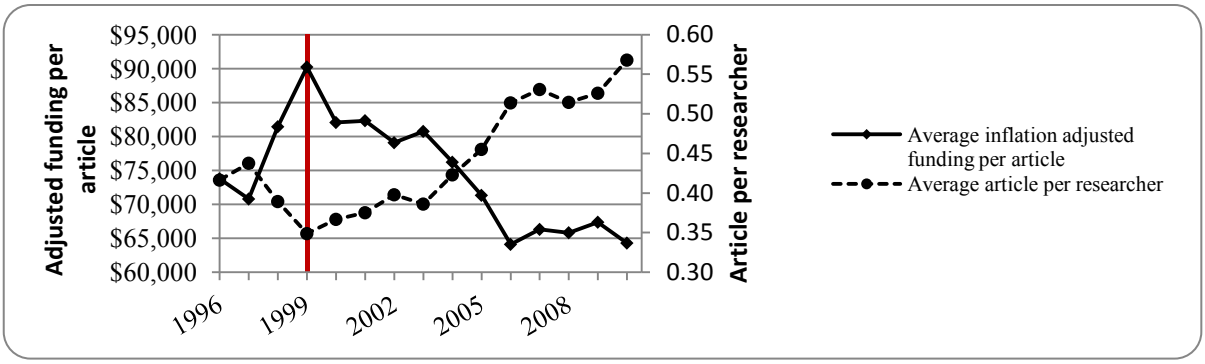


Figure 40. Average inflation adjusted funding vs. average number of articles per researcher, 1996-2010

In the rest of this section, RapidMiner software is used to apply the visualization techniques. In the following figures of this section, number of articles per year (Y-axis) and total funding per year (X-axis) are normalized to a value between 0 and 1. As expected, a considerable share of articles and funding belongs to Ontario, Quebec, British Columbia, and Alberta (Figure 41). We divided the funded researchers into three categories (junior, middle, and senior) based on their career age defined in section 5.2.1.2. In Figure 41, size of the circles represents the career age. As it can be seen, interestingly, it seems that not only the researchers from the mentioned provinces have been more productive but also the senior researchers are more located in these provinces.

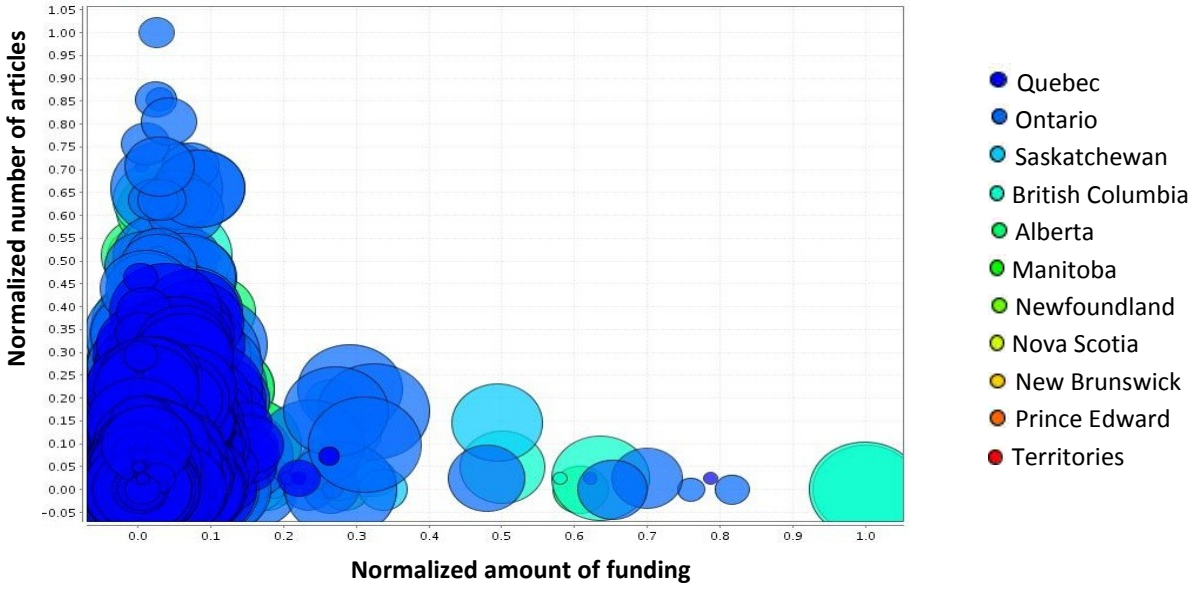


Figure 41. Funding vs. number of articles in Canadian provinces according to the career age of the researchers as circle sizes, 1996-2010

The career age of the researchers is used as the control variable. Figure 42-a shows the interaction of the career age variable with the number of articles. The number of articles was normalized to a value between 0 and 1. We considered two cases, where one is including all the funded researchers while we excluded the students in the second case. As it can be seen both curves have exactly the same trend that indicates a positive relation between age of the researchers and their productivity (except for the first and last data points in the figure). In other words, it seems that as the career age of the researcher grows his/her productivity also increases and peaks at a certain age which is highly dependent on the discipline. This finding is in line with Lehman (1953) and Lee and Bozeman (2005). In addition, the curves imply non-linear effects for which we will consider a quadratic variable in our regression. We added the funding data to the analysis which is represented in Figure 42-b as the size of the circles. Excluding the first and last data points, the figure is partially confirming that there is a positive relation between age and funding since as we move forward along the x-axis the size of the circles becomes bigger. Hence, from the visualizations it can be said that career age of the researchers and amount of funding allocated to them have a positive impact on their number of publications.

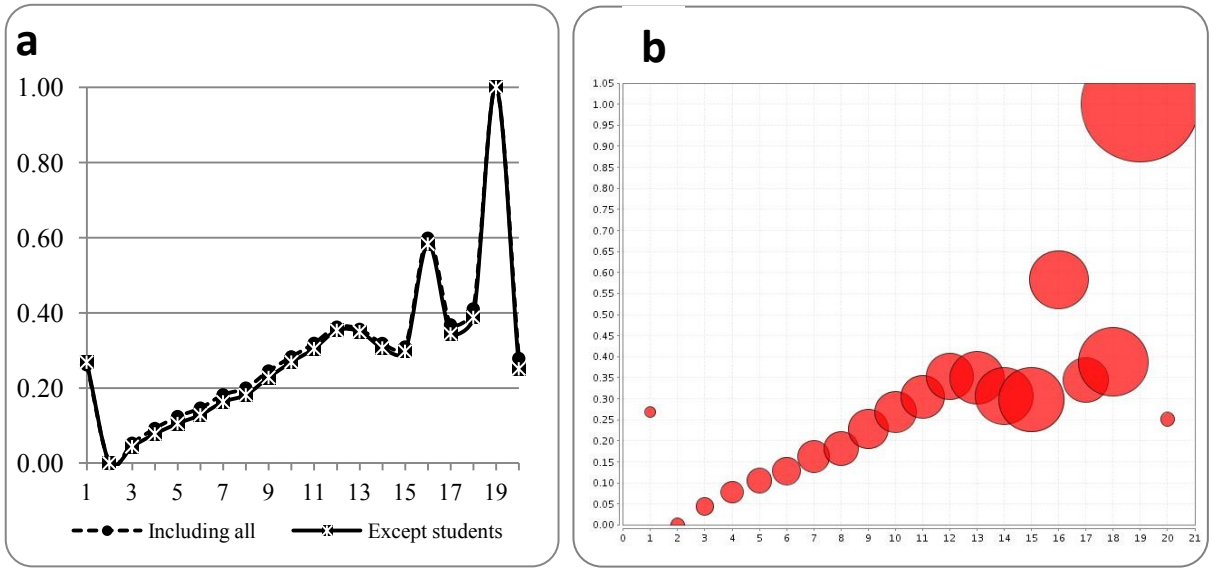


Figure 42. a) Career age vs. normalized number of publications, b) Career age, normalized number of publications, and funding as circle sizes

5.2.1.3.2 Quantity of Publications

5.2.1.3.2.1 Quantity of Publications, Complete Model

Before running the regression model, we first analyze the associations between dependent and independent variables. 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). According to Table 7, the absolute value of all the correlation coefficients is lower than 0.37, which indicates that the degree of linear correlation among the selected variables is very weak.

Table 7. Correlation matrix, complete quantity model

Variable	$noArt_i$	$ln\_avgFund3_{i-1}$	ln_avgIf3_{i-1}	$noArt1_{i-1}$	$avgTeamSize_i$	$careerAge_i$
$noArt_i$	1.0000					
$ln\_avgFund3_{i-1}$	0.2891	1.0000				
ln_avgIf3_{i-1}	0.0603	0.0804	1.0000			
$noArt1_{i-1}$	0.3655	0.2462	0.0706	1.0000		
$avgTeamSize_i$	0.0804	0.0032	0.0780	0.0346	1.0000	
$careerAge_i$	0.1577	0.3540	0.0409	0.1987	-0.0026	1.0000

Apart from the explanation given in section 5.2.1.2 in regard to the use of negative binomial predictor for our model, we also tested Poisson model and found that Poisson model does not fit our data because the goodness of fit chi-squared test was statistically significant. Hence, we employed negative binomial regression on our data to estimate the impact of the considered factors on the scientific productivity of the funded researchers measured by the number of articles in a given year. We estimated two regression models, one including all the funded researchers named as the *complete model* in the rest of the paper, while in the other model we excluded students from the data. Table 8 shows the result of the regression including all the independent, interaction, and dummy variables.

As it can be seen the average amount of researcher's funding in the past years has a significant and relatively high positive impact on the scientific production of the researcher. This is in accordance with several studies (*e.g.* Arora & Gambardella, 1998; Godin, 2003; Beaudry & Alloui, 2012) who found that larger amount of funding will result in higher number of published papers. We used the average impact factor of the journals in which a researcher published his/her articles in the past three years as a proxy for the quality of his/her work (*avgIf3*). As it can be observed in Table 8, higher quality of the papers of a researcher in the past three years increases the number of published articles. This was quite

expected since researchers who published in higher quality journals can have in general higher reputation. Higher reputation can bring higher amount of funding that might enable the researcher to expand his/her activities through finding new partners, working on new projects, *etc.* in an aim to increase the overall productivity. This relation is also confirmed by the positive overall impact of the career age of the researchers that will be discussed later.

Table 8 Negative binomial regression, the complete model

noArt	***Coef.***	***Std. Err.***	***z***	***P>|z|***	***[95% Conf. Interval]***	
ln_avgFund3	.2490258***	.0046262	53.83	0.000	.2399587	.2580929
ln_avgIf3	.0330018***	.0061728	5.35	0.000	.0209032	.0451003
noArt1	.0378271***	.0004333	87.29	0.000	.0369778	.0386764
avgTeamSize	.0026179***	.0001516	17.27	0.000	.0023208	.002915
careerAge	-.0297705***	.004317	-6.90	0.000	-.0382317	-.0213093
$careerAge^2$	.0023998***	.0002767	8.64	0.000	.0018474	.0029322
Interaction variables						
teamXage	-.000142***	.0000169	-8.40	0.000	-.0001752	-.0001089
Affiliations dummy variable						
dAcademia	.2486067***	.030134	8.25	0.000	.1895452	.3076683
Provinces dummy variables						
dQuebec	-.0794855***	.0108013	-7.36	0.000	-.1006557	-.0583154
dBColumbia	-.0475211***	.0127599	-3.72	0.000	-.0725299	-.0225122
dAlberta	.0765243***	.0132953	5.76	0.000	.0504661	.1025826
dSaskatchewan	.005897	.0235165	0.25	0.802	-.0401944	.0519884
dNBrunswick	-.087087***	.0302696	-2.88	0.004	-.1464143	-.0277598
dManitoba	.0101386	.0241655	0.42	0.675	-.037225	.0575022
dNFoundland	-.0472348	.0315774	-1.50	0.135	-.1091253	.0146558
dPEdward	-.0771565	.0792196	-0.97	0.330	-.232424	.078111
dNScotia	-.0693328***	.021324	-3.25	0.001	-.111127	-.0275386
Funding programs dummy variables						
dStrategic	.0838271***	.0176068	4.76	0.000	.0493183	.1183359
dTools	.0281414	.0249109	1.13	0.259	-.0206831	.076966
dCollaborative	-.0115878	.0175147	-0.66	0.508	-.0459161	.0227404
dIndustrial	.0680523**	.026957	2.52	0.012	.0152176	.1208871
dStudent	-.2530735***	.0188543	-13.42	0.000	-.2900273	-.2161197
dOther	-.004898	.011799	-0.42	0.678	-.0280236	.0182275
_cons	-2.85391***	.0536381	-53.21	0.000	-2.959038	-2.748781
ln(alpha)	-.6738737***	.013409			-.700155	-.6475925
alpha	.5097302	.006835			.4965084	.5233041
Likelihood-ratio test of alpha=0:			chibar2(01) = 1.5e+04		Prob>=chibar2 = 0.000	

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 84,048

According to the results, past productivity (*noArt1*) of researchers has also a positive effect on their number of publications. This is also expected since it is more probable that a

researcher with higher productivity attracts more funds that might result in higher number of publications. In addition, it is more likely that a productive researcher at least maintains his/her level of productivity in the coming year. Moreover, according to the results the average team size of the researchers (*avgTeamSize*) positively influences their productivity. Larger scientific team size can enable researchers to better distribute the work among the team members. It would be also possible to work on larger or more projects. Hence, in general we can assume that members of the larger teams have better access to scientific resources (*e.g.* expertise, equipments, and finance) which will help them to increase the scientific productivity. Although there are also some disadvantage of having larger team (*e.g.* coordination costs), according to our dataset the overall impact of team size on the productivity of the NSERC funded researchers is positive.

The career age of the funded researchers that we employed as the control variable has an overall positive impact on the number of publications. We first considered the model without the quadratic term that resulted in a positive coefficient for the career age variable (0.0062546). As explained in Figure 42-a, non-linear effects were observed for the career age of the researchers. We added the quadratic term to see the curvature of the relationship. Hence, the predictive effect of the researchers' career age is represented by $\beta_1 careerAge + \beta_2 careerAge^2$ which is increasing over the range of the career age. Therefore, number of publications increases with the career age of the researchers. From the regression results and the discussion it seems that our dataset partially verifies the existence of the Matthew effect (Merton, 1968) in a sense that the circle of higher quantity and quality of the past works, higher reputation, and higher amount of money available attracts more money to a researcher in a way that the rich gets richer. Although both career age and team size have positive impact on the number of publications, interestingly the interaction variable has a negative and significant effect. This may imply that as the career age of the researchers increases, larger team sizes can affect their number of publications negatively. In other words, large team sizes can negatively influence the productivity of aged funded researchers.

In order to dig further into where the NSERC funding has had a stronger effect in terms of the number of publications, we included dummy variables in the regression representing the institution type and Canadian provinces. We also considered dummy variables for

NSERC funding programs to compare the impact of the programs. The institution type dummy variable (*dAcademia*) takes value 1 if the funded researcher is affiliated to the academic institution and 0 if his affiliation is non-academic. According to Table 8, academic funded researchers are significantly different from the non-academic ones and are producing around 25% (0.249) more than the non-academic researchers. Analysis of the provinces dummy variables reveals that the funded researchers from Quebec, British Columbia, Alberta, New Brunswick, and Nova Scotia are significantly different from the ones who reside in Ontario which was the omitted dummy variable. Interestingly, among the mentioned provinces only the coefficient of Alberta dummy variable is positive (0.0765243) which shows higher productivity of Alberta's funded researchers.

To analyze the effect of different NSERC funding programs, we categorized the programs into seven categories: Discovery grants, strategic projects, collaborative grants, student grants, tools, industry grants, and other programs. We considered the discovery grants as the omitted variable. From the analysis it can be seen that the effects of strategic, industrial, and student grants are significantly different from the discovery grants program while the effect is only negative for the student grants (-.2530735). According to the definition of these grants the results are quite as expected. Specifically for the strategic project grants, the aim is to improve the scientific development in selected high-priority areas that influences Canada's economic and societal position. Hence, these narrowly-defined targeted grants should be allocated to specific reputable researchers who are more productive. In the next section we remove the students' data and analyze the model for the rest of the funded researchers.

5.2.1.3.2.2 Quantity of Publications, Students-Excluded Model

We removed the students' data and performed the regression for the rest of the researchers. For this purpose, we labeled a researcher "*student*" in a year whenever his/her highest average grant was coming from one of the student funding programs in that year. Moreover, to better account for the quality of the work of the funded researchers we also considered the average number of citations of their articles in the past three years ($avgCit3_{i-1}$). The correlation matrix presented in Table 9 shows a weak linear correlation degree among the considered variables.

Table 9. Correlation matrix, student-excluded quantity model

Variable	$noArt_i$	$ln\_avgFund3_{i-1}$	ln_avgIf3_{i-1}	$avgCit3_{i-1}$	$noArt1_{i-1}$	$avgTeamSize_i$	$careerAge_i$
$noArt_i$	1.0000						
$ln\_avgFund3_{i-1}$	0.2984	1.0000					
ln_avgIf3_{i-1}	0.0305	0.1072	1.0000				
$ln\_avgCit3_{i-1}$	0.1248	0.1434	0.4084	1.0000			
ln_noArt1_{i-1}	0.4478	0.2799	0.0328	0.0249	1.0000		
$avgTeamSize_i$	0.0856	0.0269	0.0982	0.0665	0.0488	1.0000	
$careerAge_i$	0.1602	0.3378	0.0193	0.2495	0.1787	-0.0058	1.0000

The results of the negative binomial regression for the student-excluded model of the number of publications are shown in Table 10. As it can be seen, average journal impact factor (*avgIf3*) has a significant negative impact on the quantity of the publications, while average citations (*avgCit3*) have a positive effect. The intensity of the both mentioned factors is almost the same. Hence, it can be said that the quality of the papers of the pure scientists (students excluded) measured by the average number of citations in the past three years influences the number of publications positively. The citation-based proxy seems to be a better measure for evaluating the quality of the pure researchers' papers. According to the regression results, it can be said that the researchers with high amounts of funding publish relatively low quality papers in high quality journals. These papers would not be highly cited, which justifies the negative coefficient of *avgIf3*.

Other interesting finding is the high impact of a researcher's past productivity (*noArt1*) on the number of publications. Hence, not only the quality of the works in the past plays an important role in higher productivity but also the rate of the publications is a major sign of productive researchers. The career age of the researchers is also showing a positive impact while the quadratic term ($careerAge^2$) affects negatively. Hence according to the curvature of the relationship, although our study covers 15 years from 1996 to 2010, it can be predicted that around 18 years after the start of the work of a NSERC funded researcher[17] his/her scientific productivity starts to decline. Therefore, mid-career NSERC funded researchers seem to be more productive. This finding is in line with Cole (1979), Wray (2003), Wray (2004), Kyvik and Olsen (2008), and Beaudry and Allaoui (2012) who also found the higher scientific productivity of the mid-career aged researchers.

[17] Measured by the date of his first publication that is available on Scopus.

Table 10. Negative binomial regression, student-excluded (pure) model

noArt	*Coef.*	*Std. Err.*	*z*	*P>|z|*	*[95% Conf. Interval]*	
ln_avgFund3	.1948294***	.0056854	34.27	0.000	.1836861	.2059726
ln_avgIf3	-.1066736***	.0091611	-11.64	0.000	-.124629	-.0887182
ln_avgCit3	.1163277***	.0053229	21.85	0.000	.105895	.1267603
ln_noArt1	.5851529***	.0071044	82.36	0.000	.5712285	.5990772
avgTeamSize	.0022361***	.0002179	10.26	0.000	.0018089	.0026632
careerAge	.066107***	.005784	11.43	0.000	.0547707	.0774434
careerAge²	-.0038581***	.0003453	-11.17	0.000	-.0045349	-.0031813
Interaction variables						
teamXage	-.0000895***	.0000239	-3.74	0.000	-.0001364	-.0000426
Affiliations dummy variable						
dAcademia	.2690371***	.0377299	7.13	0.000	.195088	.3429863
Provinces dummy variables						
dQuebec	-.058617***	.0127778	-4.59	0.000	-.083661	-.033573
dBColumbia	-.0349328**	.0150169	-2.33	0.020	-.0643653	-.0055003
dAlberta	.0656808***	.015411	4.26	0.000	.0354758	.0958857
dSaskatchewan	.0126419	.0273714	0.46	0.644	-.0410051	.0662888
dNBrunswick	-.0552766	.0363791	-1.52	0.129	-.1265783	.0160251
dManitoba	.0250575	.0283492	0.88	0.377	-.030506	.0806209
dNFoundland	-.0408263	.039075	-1.04	0.296	-.1174118	.0357592
dPEdward	-.0479439	.0933412	-0.51	0.608	-.2308893	.1350015
dNScotia	-.0525115**	.0250681	-2.09	0.036	-1016441	-.0033788
Funding programs dummy variables						
dStrategic	.0642019***	.0195962	3.28	0.001	.025794	.1026098
dTools	.0047521	.0287229	0.17	0.869	-.0515438	.0610479
dCollaborative	-.0310454	.0200215	-1.55	0.121	-.0702868	.008196
dIndustrial	.0729073**	.0302634	2.41	0.016	.0135921	.1322224
dOther	-.0377083***	.0135293	-2.79	0.005	-.0642253	-.0111914
_cons	-2.535194***	.0658431	-38.50	0.000	-2.664244	-2.406144
ln(alpha)	-1.058564***	.019625			-1.097028	-1.0201
alpha	.3469538	.006809			.3338618	.3605591
Likelihood-ratio test of alpha=0:			chibar2(01) = 6349.65		Prob>=chibar2 = 0.000	

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 43,514

Other estimated factors including the dummy variables are showing the same effect as the ones that predicted by the complete model in section 5.2.1.3.2.1. The only exception is for the dummy variable of New Brunswick province that becomes no longer significant in the students-excluded model indicating that there is no significant difference between the researchers of New Brunswick and the omitted province of Ontario. In the next section, we discuss the impact of the influencing factors on the quality of the researchers' papers.

5.2.1.3.3 Quality of Publications

5.2.1.3.3.1 Quality of Publications, Complete Model

In this section, impact of the influencing factors on quality of the works of the researchers is investigated. The impact is assessed for all the researchers in the database, including the students. The correlation matrix of the considered variables is presented in Table 11, which reports a very weak linear correlation for most of the variables. The absolute value of the correlation coefficients is less than 0.4.

Table 11. Correlation matrix, complete quality model

Variable	$noArt_i$	$ln\_avgFund3_{i-1}$	$avgArt3_{i-1}$	$avgIf3_{i-1}$	$avgCit3_{i-1}$	$avgTeamSize_i$	$careerAge_i$
$avgCit_i$	1.0000						
$ln\_avgFund3_{i-1}$	0.0536	1.0000					
$avgArt3_{i-1}$	0.0833	0.3965	1.0000				
$avgIf3_{i-1}$	0.1394	0.1198	0.1910	1.0000			
$avgCit3_{i-1}$	0.1845	0.0680	0.1002	0.3028	1.0000		
$avgTeamSize_i$	0.0614	0.0016	0.0197	0.0551	0.0307	1.0000	
$careerAge_i$	0.0111	0.3548	0.3630	0.0802	0.0986	-0.0114	1.0000

Since in the quality of papers model the dependent variable (*avgCit*) is not a count measure we used the multiple regression analysis for estimating the impact of the considered factors on the quality of the papers of the NSERC funded researchers. According to Table 12, all the independent variables significantly influence the quality of the papers measured by average number of citations. As expected, past funding (*avgFund3*) has a positive impact on the quality of the papers. This is interesting since in the literature mainly no relation is found between funding and quality of the works (*e.g.* Godin, 2003; Payne & Siow, 2003).

The past productivity (*avgArt3*) and the quality of the past works of the funded researchers also affect positively the average citations received by their papers in the current year. Hence, this is implicitly confirming that productive researchers with high quality previous works may continue producing high quality papers. As expected, researchers who get involved in larger scientific teams also produce higher quality papers since they can benefit from internal referring among the team members that can improve the quality of the paper. Another interesting point is the negative relation observed between the career age of the funded researchers and the quality of their work. Hence, the results suggest that as the career age of the researcher increases they produces on average lower quality papers. This

can be caused by several factors, *e.g.* lower motivation, or higher reputation in a way that papers are published but not necessarily highly cited, *etc*. This supports the finding from the previous model where we observed that highly funded researchers on average publish in high ranking journals but their work is less cited. Also, as the career age of the researchers augments, larger team sizes influence the quality of papers negatively (*teamXage*).

Table 12. Regression results, complete quality model

avgCit	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf. Interval]*	
ln_avgFund3	.3537779***	.0245286	14.42	0.000	.3057021	.4018537
avgArt3	.2763136***	.0170422	16.21	0.000	.2429111	.309716
avgIf3	.4535849***	.0189189	23.98	0.000	.4165041	.4906657
avgCit3	.1170045***	.0024088	48.57	0.000	.1122833	.1217258
avgTeamSize	.008713***	.0005863	14.86	0.000	.0075639	.009862
careerAge	-.1470174***	.0214365	-6.86	0.000	-.1890326	-.1050023
careerAge^2	.0084294***	.0014472	5.82	0.000	.0055928	.011266
Interaction variables						
teamXage	-.000511***	.0000798	-6.40	0.000	-.0006674	-.0003546
Affiliations dummy variable						
dAcademia	-.6036918***	.1373958	-4.39	0.000	-.8728855	-.3342981
Provinces dummy variables						
dQuebec	-.1449621**	.0571347	-2.54	0.011	-.2569452	-.032979
dBColumbia	.3221105***	.0682374	4.72	0.000	.1883662	.4558549
dAlberta	-.3136267***	.0737739	-4.25	0.000	-.4582226	-.1690309
dSaskatchewan	-.4159663***	.1246093	-3.34	0.001	-.6601986	-1717339
dNBrunswick	-.7473366***	.1499116	-4.99	0.000	-1.041161	-.4535121
dManitoba	-.6972798***	.1289009	-5.41	0.000	-.9499237	-.4446359
dNFoundland	-.6921608***	.1653934	-4.18	0.000	-1.016329	-.3679923
dPEdward	1.723699***	.4082086	4.22	0.000	.9236167	2.523782
dNScotia	-.2261751**	.1106283	-2.04	0.041	-.443005	-.0093452
Funding programs dummy variables						
dStrategic	.2671592***	.1027175	2.60	0.009	.0658344	.4684839
dTools	.3709673***	.1403731	2.64	0.008	.095838	.6460966
dCollaborative	.0247293	.0964148	0.26	0.798	-.1642423	.2137008
dIndustrial	.1869615	.1425337	1.31	0.190	-.0924023	.4663254
dStudent	2.069167***	.0765411	27.03	0.000	1.919148	2.219187
dOther	1.059779***	.0652033	16.25	0.000	.9319813	1.187576
_cons	-.5555149**	.270778	-2.05	0.040	-1.086236	-.0247941

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 111,994

Analyzing dummy variable of the institution type reveals that the funded researchers who are affiliated with the industry are producing on average higher quality papers measured by the average number of citations. Regarding the provinces, all the Canadian provinces dummy variables are significantly different from Ontario which is the omitted dummy

variable. The coefficient is negative for all the provinces except for British Columbia and Prince Edward. However, nothing can be concluded about the funded researchers located in Prince Edward province since the number of articles, number of researchers, and the total amount of funding is much lower there in comparison with other provinces. We omitted the discovery grants dummy variable for analyzing the impact for different NSERC funding programs. As it can be seen, *dStrategic*, *dTools*, *dStudent*, and *dOther* are significantly and positively different from the omitted program. This finding was expected for the Strategic funding programs but not expected for the student programs. In general, it can be said that more limited scope of a funding program with more narrowly defined targets can result in higher quality papers. On the other hand, one may not expect a direct positive impact of very general programs (*e.g.* discovery grants) since they cover almost all the funded researchers.

5.2.1.3.3.2 Quality of Publications, Student-Excluded Model

In this section, the same variables and analysis are used on the student-excluded data. Table 13 reports the linear correlations among the considered variables. The absolute value of the correlation coefficients is less than 0.38, which is the correlation between the past average productivity (*avgArt3*) and past average funding (*avgFund3*). We continue with the multiple regression analysis on the data since the correlations are not significant.

Table 13. Correlation matrix, student-excluded (pure) quality model

Variable	$noArt_i$	$ln\_avgFund3_{i-1}$	$avgArt3_{i-1}$	$avgIf3_{i-1}$	$avgCit3_{i-1}$	$avgTeamSize_i$	$careerAge_i$
$avgCit_i$	1.0000						
$ln\_avgFund3_{i-1}$	0.0998	1.0000					
$avgArt3_{i-1}$	0.1078	0.3775	1.0000				
$avgIf3_{i-1}$	0.1601	0.1175	0.1885	1.0000			
$avgCit3_{i-1}$	0.2149	0.0792	0.1047	0.3035	1.0000		
$avgTeamSize_i$	0.0627	0.0326	0.0237	0.0542	0.0329	1.0000	
$careerAge_i$	0.0454	0.2842	0.3381	0.0706	0.1072	-0.0033	1.0000

Table 14 shows the regression results for the student-excluded quality model. The sign of the resulted variables are exactly the same as the ones for the complete quality model while the coefficients are also almost the same. Hence, the justifications that were presented in section *5.2.1.3.3.1* hold. The only difference is for the career age variable (*careerAge*) and industrial programs dummy variable (*dIndustrial*). According to Table 14, the dummy variable for the industrial funding programs is showing significantly different impact (with the coefficient of 0.26) in comparison with the omitted dummy variable of the discovery grants. The career age of the NSERC funded researchers in the student-excluded model has

become insignificant indicating that the career age of the NSERC funded researcher does not make a difference on the quality of the papers produced. This is quite interesting since it shows that the career age does not affect a researcher to produce a high quality paper, whereas other factors like the amount of funding, past productivity, quality of the past papers are playing a more important role in this regard. Therefore, it can be proposed that more equal distribution of NSERC funding among the young and senior researchers who possess a good scientific profile can result in higher quality papers.

Table 14. Regression results, student-excluded quality model

avgCit	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf.*	*Interval]*
ln_avgFund3	.3490946***	.024293	14.37	0.000	.3014807	.3967086
avgArt3	.25321***	.0158916	15.93	0.000	.2220627	.2843574
avgIf3	.4613971***	.0189419	24.36	0.000	.4242712	.498523
avgCit3	.1299262***	.0024331	53.40	0.000	.1251573	.134695
avgTeamSize	.0113608***	.0007585	14.98	0.000	.0098742	.0128474
careerAge	-.0105686	.0205416	-0.51	0.607	-.0508299	.0296926
careerAge^2	.0004639	.0013778	0.34	0.736	-.0022366	.0031645
Interaction variables						
teamXage	-.0007191***	.000088	-8.17	0.000	-.0008916	-.0005465
Affiliations dummy variable						
dAcademia	-.576718***	.129011	-4.47	0.000	-.8295779	-.323858
Provinces dummy variables						
dQuebec	-.1447395***	.0555324	-2.61	0.009	-.2535824	-.0358967
dBColumbia	.3213612***	.0669523	4.80	0.000	.1901354	.4525869
dAlberta	-.3296383***	.0725502	-4.54	0.000	-.4718357	-.1874409
dSaskatchewan	-.3856221***	.1182055	-3.26	0.001	-.6173034	-.1539409
dNBrunswick	-.692624***	.1435	-4.83	0.000	-.9738822	-.4113658
dManitoba	-.5918489***	.1233531	-4.80	0.000	-.8336195	-.3500783
dNFoundland	-.6801265***	.1585457	-4.29	0.000	-.9908742	-.3693789
dPEdward	1.950506***	.3946359	4.94	0.000	1.177024	2.723987
dNScotia	-.2155348**	.1080504	-1.99	0.046	-.4273122	-.0037574
Funding programs dummy variables						
dStrategic	.2788403***	.0947333	2.94	0.003	.0931642	.4645164
dTools	.3724214***	.1292664	2.88	0.004	.1190607	.625782
dCollaborative	.0372067	.0888975	0.42	0.676	-.1370313	.2114447
dIndustrial	.2645677**	.1315162	2.01	0.044	.0067976	.5223378
dOther	1.024165***	.0612311	16.73	0.000	.9041533	1.144178
_cons	-.9746751***	.2648914	-3.68	0.000	-1.493859	-.4554912

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 99,216

5.2.1.4 Conclusion

In this paper we investigated the impact of funding and other influencing factors like scientific team size and past productivity on quantity and quality of the publications of the funded researchers. All the four regression models confirmed the significant positive impact of funding on the productivity of the researchers. The positive relation between funding and the rate of publications has been also confirmed in the work of other scholars, *e.g.* Arora & Gambardella (1998), Boyack and Borner (2003), Payne and Siow (2003), Jacob and Lefgren (2007), Zucker *et al.* (2007), and Beaudry and Allaoui (2012). However, to our knowledge the studies that used statistical analysis to assess the productivity of the funded researchers have found no impact of funding on the quality of the papers (*e.g.* Godin, 2003; Payne & Siow, 2003) where in our case we found a significant positive relation. The past productivity of a funded researcher in terms of both quantity and quality of his/her publications was also indicated as one of the important factors that positively affects the rate and quality of publications of the funded researcher in the current year. Although according to our results higher level of funding may result in higher scientific performance, since the financial resources are limited it would be proposed that NSERC give higher weights to more productive researchers regardless of their age and reputation. Of course this allocation strategy needs to be reviewed and revised annually.

The other interesting finding was in regard to the impact of the career age on productivity of a funded researcher. For the quantity of the publications model it has been observed that mid-career funded researchers seem to be more productive that is in line with the work of other scholars like Cole (1979), Wray (2003), Wray (2004), Kyvik and Olsen (2008), and Beaudry and Allaoui (2012). However, no significant effect was observed for the career age related variables in the student-excluded quality of the papers model. This may implicitly highlight the importance of more equal funding distribution among young and senior researchers who have a prolific scientific profile, especially in well targeted high priority funding NSERC programs like the strategic project programs.

We also compared the impact of different NSERC funding programs on scientific output of the funded researchers to find out which program yields the highest productivity. As expected, strategic programs which are of high priority and narrower scope showed positive

and significant impact in all the four analyzed models in comparison with the omitted dummy variable of the discovery grants. Interestingly, the provinces dummy variables were significantly different from the omitted dummy variable of Ontario in the quality model where the coefficient was positive only for the researchers located in British Columbia and Prince Edward provinces. And, analyzing the dummy variable of the institution type (*dAcademia*) reveals that although the NSERC funded researchers who were affiliated with academic institutions were more productive in terms of the number of publications, the papers of the industry affiliated funded researchers have been of higher quality.

5.2.1.5 Limitations and Future Work

We were exposed to some limitations in this paper. Scopus (and other similar databases) that was selected as the source of data is English biased, hence, non-English articles are underrepresented (Okubo, 1997). We were forced to choose 1996 as the beginning year of the analysis since Scopus data was less complete before 1996. Another inevitable limitation related to 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.

Different scientific disciplines follow different patterns in publishing articles, collaborating with other researchers, or even getting and allocating grants. Hence to better examine scientific productivity and efficiency, a future work direction could be assessing the impact of funding on the rate of publications for different scientific disciplines separately. In addition, other funding councils can be considered as the source of funding data. This kind of analyses and comparing the efficiency of different funding organizations may help the decision makers to set the best funding allocation strategy.

5.2.2 On the Impact of the Small World Structure on Scientific Activities

The modern science has become more complex and interdisciplinary in its nature which might encourage researchers to be more collaborative and get engaged in larger collaboration networks. Various aspects of collaboration networks have been examined so far to detect the most determinant factors in knowledge creation and scientific production. One of the network structures that recently attracted much theoretical attention is called *small world*. It

has been suggested that small world can improve the information transmission among the network actors. In this section, using the data on 12 periods of journal publications of Canadian researchers in natural sciences and engineering, the co-authorship networks of the researchers are created. Through measuring small world indicators, the small worldiness of the mentioned network and its effect on productivity, quality of publication, and team size are assessed. Our results show that the examined co-authorship network strictly exhibits small world properties. In addition, it is suggested that in a small world network researchers expand their team size through getting connected to other experts of the field. This team size expansion may result in higher productivity of the whole team as a result of getting access to new resources, benefitting from the internal referring, and exchanging ideas among the team members. Moreover, although small world network has a positive impact on the quality of the articles in terms of both the number of received citations and journal impact factors, it negatively affects the average productivity of researchers in terms of the number of their publications.

5.2.2.1 Introduction

The world is really small! This comes to our minds when we find a mutual acquaintance with someone who we do not know at all. The idea of the small world network is traced back to the work of Milgram in 1967. Through a series of field experiments he found that even in a very large network on average only six intermediates are needed to reach a person who is completely unknown[18]. This property is also called "*six degrees of separation*" in the literature (Guare, 1992). In other words, in the small world networks the average path length[19] is relatively short in spite of the existence of high clustering[20]. Therefore, short path lengths among network actors facilitate the spread of various ideas that are generated in separate clusters, which results in producing novel knowledge (Uzzi & Spiro, 2005; Fleming & Marx, 2006).

The level and the efficiency of knowledge diffusion are affected by small world property. Cowan and Jonard (2004) developed a model to study the efficiency of small world

[18] Later, Travers and Milgram (1969) tried to formulate the small world property by calculating the probability of any two randomly chosen people knowing each other in a large population.

[19] Average distance between two given nodes in the network.

[20] Tendency of the nodes in a network to cluster together.

networks and claimed that the level of knowledge is at its maximum when the network structure has small world properties. Therefore, it is good to have small world property in the network but how persistent are such networks? Kogut and Walker (2001) analyzed the cross-ownership among German firms during 1990s and the robustness of the small world property. They found that the small world network tends to preserve its properties of high clustering and short path lengths even if it experiences a considerable number of shocks and re-structuring of the links of the network. Therefore, once the small world network is established it retains the property unless the network perceives a considerable amount of re-structuring forcing it to transform into another structure.

Several researchers analyzed the effect of small world property in the network of firms. Sullivan and Tang (2012) constructed the inter-firm network of the United States venture capital industry to evaluate its effects on the firms' performance. They observed a positive impact of small world structure on productivity of firms. In another study, Kogut and Walker (2001) investigated the Canadian network of investment bank syndicate from 1952 to 1990 to see how small world network emerges and evolves over time. They confirmed that the networks formed among firms usually resemble small world characteristics. Schilling and Phelps (2007) focused on the impact of the small world property on firms' performance through analyzing the number of patents. Their results show that there is a positive effect of the small world since high clustering and short path length enables companies to get access to new knowledge that is required for innovation.

In addition, several empirical studies focused on individuals' activity and analyzed the effect of the small world property on the performance of individuals in the network. Fleming and Marx (2006) studied the collaboration of the inventors in Silicon Valley and Route 128 in Boston and found that the network of the examined inventors resembled the small world structure. However, no positive relation was observed between the existing small world property in the network and the inventive productivity of the researchers in the region. Fleming *et al.* (2007) have also shed some light on the impact of the small world on the network of inventors and their innovative and managerial approaches within a small world network to remain competitive. Although they found a positive effect of short average path

length on the technological productivity, no significant positive influence of the small world property was observed.

Other studies analyzed the impact of the small world structure in co-authorship networks. Co-authorship analysis has been particularly recognized by some studies (*e.g.* Glanzel, 2001; Savanur & Srikanth, 2010) as being the most common tool in investigating the relations and patterns in scientific collaboration. Newman (2004) investigated the co-authorship networks in physics, biology and mathematics and found the small world structure in all the aforementioned networks. Goyal *et al.* (2006) focused on a single scientific discipline. Using the co-authorship network of economists during 1980 to 1999, they found small world properties in the examined collaboration network. Moreover, they found an increasing trend in the average degree of the network over time and realized that the number of brokers is also augmenting. In another study, despite considering several fields for the study Moody (2004) also focused on the subspecialties (*e.g.* economic sociology, criminology, *etc*.) in a single discipline and analyzed the network of sociologists during the period of 1963 to 1999. He surprisingly found that the network did not resemble the small world properties likely due to the considerable overlap among the subfields and the authors.

Hence, there is a tendency in co-authorship networks for the small world structure. Role of the best connected actors in joining the other individuals and clusters in the network is very important. Moreover, the co-authorship pattern in a scientific field is also crucial for a network to obtain small world structure. The more a scientific discipline is team oriented and the larger the size of the team, the more probability of finding the small world properties in the structure (Guimerà, *et al.*, 2005; Wuchty, *et al.*, 2007). Therefore, the analysis of small world property is more seen in the disciplines in which teamwork is common (Lissoni, *et al.*, 2013).

Studies that have generally assessed the impact of network structure variables in co-authorship networks have found correlations between the centrality measures and some performance variables (Yan & Ding, 2009; Abbasi, *et al.*, 2011; Kumar & Jan, 2013; Eslami, *et al.*, 2013). Yan and Ding (2009) focused on 16 journals in the field of library and information science (LIS) and constructed the co-authorship network at the micro level over

the time span of 1988 to 2007. They calculated four centrality measures for the authors in the network, *i.e.* betweenness centrality, degree centrality, closeness centrality and PageRank and found a positive relation between the mentioned measures and citation counts of articles. Abbasi *et al.* (2011) focused on the scholars in the field of information systems and statistically analyzed the impact of the network structure variables on the performance of the researchers using citation based indicator. They found a positive relation between all the network structure variables and the performance of the scholars except for the betweenness and closeness centralities. In another study, Kumar and Jan (2014) assessed and compared the impact of the network variables in the field of energy fuels on research performance in Turkey and Malaysia. According to their results, popularity, position and prestige of the researchers measured by the network centrality indicators have a positive impact on their research performance. In addition, they found PageRank as the most influential centrality measure. Eslami *et al.* (2013) focused on the field of biotechnology in Canada and statistically investigated the impact of the network structural variables on the quantity and quality of technological performance of the researchers within the period of 1966 to 2005. Their results suggest a significant impact of structure of the examined co-authorship network on knowledge and technology production, however, no impact was observed on the quality of the patents.

Nevertheless, the results about the impact of the small world structure on performance are inconsistent. For example, Fowler (2005) found a non-linear relation between small world properties and voting participation rate, and Uzzi and Spiro (2005) found a similar relation between the financial and artistic performance of the artists and the small world properties. However, Schilling and Phelps (2007) observed a linear relation whereas Fleming *et al.* (2007) found no relation between small world properties and performance. Hence, no consensus is found in the literature about the impact of the small world structure on the performance (Uzzi, *et al.*, 2007). One reason could be the use of different datasets and performance measures in the studies that makes it hard to come into a general agreement about the impact of the small worldiness on researchers' performance. Hence, the assessment of the impact is suggested to be done in different fields and scientific environments. In addition, although there are very few studies that particularly analyzed the impact of the small world variables on productivity of the inventors and firms, to our knowledge no study

has analyzed such impact on the quality of the publications and researchers' team size. This paper is designed to fill these research gaps.

Our main objective is to study the impact of the small world network structure on the scientific output, on the quality of the produced papers and on the team size. It is assumed that analyzing the impact of small world property on the quality of the publications will help to highlight the benefits of a systematic collaboration network rather than a random one in producing higher quality research. In addition, it will indentify the importance of a well-established collaboration network in which researchers are well connected by short distances. Moreover, analyzing the impact of the small world property on the average team size of the researchers will determine if researchers in a small world network prefer to have larger team sizes due to the shorter distance among researchers in such a network. As larger team size may result in higher rate of publication, if the impact of the small world property is positive on the team size then one may expect higher rate of publications in such collaboration networks.

In order to achieve this objective we use a comprehensive dataset of the publications of Canadian researchers in natural sciences and engineering. First we examine the existence of the small world properties in the co-authorship network of these researchers and then statistically investigate the effects of the small world variables on the quantity of the scientific output (measured by the number of publications), quality of the articles (measured by the normalized citation rate and by the average impact factor of the journals) and on the size of the research teams (represented by the average number of authors per paper). The rest of the paper is organized as follows: Section 5.2.2.2 describes methodology and data that was used in this study. The empirical results and interpretations are provided in section 5.2.2.3. Section 5.2.2.4 presents the findings of this research and the limitations of this study are discussed in the last section 5.2.2.5.

5.2.2.2 Data and Methodology

The study has three phases. In the first phase, we created a database of all the research publications produced by the Canadian researchers in natural sciences and engineering. We decided to focus only on engineering and natural sciences and to exclude social and medical

sciences, because collaboration patterns in different disciplines vary[21]. In order to do so we included only the researchers funded by Natural Sciences and Engineering Research Council (NSERC), which is the main Canadian federal funding agency for the researchers working in all the areas of engineering and natural sciences. Since almost all the Canadian researchers in these research fields are currently receiving or received in the past a research grant from NSERC (Godin, 2003), we assumed that this approach will allow us to identify them quite effectively. We found this procedure more straightforward than collecting all the Canadian papers and trying to distinguish between the ones that are written by the researchers in natural sciences and engineering and other scientific fields through employing some keywords or journal categories. Eligibility for NSERC funding makes our target researchers clearly defined.

Then we collected from Scopus the articles written by these researchers in the period of 1996 to 2010 since the data quality of Scopus was low before 1996. Moreover, 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 in. SCImago was chosen for two main reasons. Firstly, it provides annual data of the journal impact factors that enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our articles database. In total, the final database contained 130,510 articles and 177,449 authors together with all the related information (*e.g.* article title, co-authors, their affiliations, year of publication).

In the second phase, we used Pajek software[22] to construct the collaboration networks of the researchers and to measure the structural network and small world variables. Co-authoring an article was assumed as a sign of collaboration among the researchers, but we had no information on the length of this relationship. In some of the similar studies (*e.g.* Baum, *et al.*, 2003; Fleming, *et al.*, 2007) a 5-year period for the life of each created collaboration link in the networks was considered while in other studies a 3-year time window is assumed (*e.g.* Beaudry & Allaoui, 2012). We calculated the indicators for both of the mentioned time windows and found that the results are more robust for the 3-year time

[21] As an example, please see Larivière *et. al* (2006).
[22] For more information, see: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

window. Hence, we assumed a 3-year time window in our study and shifted the 3-year moving window forward from 1996 to 2010 to extract the publications for each of the networks. This procedure resulted in 12 undirected networks. The structure of the 12 networks was then analyzed separately by Pajek software to measure the small world variables for each of the 12 networks.

In the last phase, the measures calculated in the previous phase were used as inputs to the models to statistically analyze the impact of small world properties on the productivity and scientific collaboration of the scientists. For this purpose, five regression models were defined and estimated by STATA software. The first dependent variable accounts for the research productivity of the researchers within each of the 12 periods (*no_art*). The number of publications has been widely used in the literature as the quantity proxy of scientific productivity (*e.g.* Centra, 1983; Okubo, 1997). We considered a single year for representing the productivity of the researchers since we assumed that the results of researchers' collaboration come to light soon after the respective collaboration period is finished (as was done in Baum, *et al.*, 2003; Fleming, *et al.*, 2007). In other words, it is assumed that the 3-year collaborative activity among the researchers will be reflected in the next year in the form of the number of their publications. Hence, for the total number of articles in the year *i* (no_art_i), we calculated the small world variables for the networks constructed on the 3-year snapshot from year *i-3* to *i-1*. In order to investigate the impact on productivity more accurately, we normalized the number of publications by dividing them by the number of authors and considered it as the dependent variable for the second regression model ($art\_per\_aut_i$). This may help us to better analyze the direct impact of the small world variables on productivity since higher number of authors may result in higher number of publications. Hence, by averaging the number of publications over the number of co-authors the impact of the raise in the number of authors will be accounted. In order to assess the quality of the publications we used the normalized number of citations in the third model. Citation count based indicators are one of the most widely used approaches in determining research quality (Kostoff, 2002). However, like all the methods it has some drawbacks, *e.g.* negative citations, self citations (Okubo, 1997), and limitations of the citation data source (Couto, *et al.*, 2009). Nevertheless, it is generally accepted in bibliometrics that the real or expected number of citations received by publications can be used as a good index of the

mean impact at the aggregate level (Seglen, 1992; Gingras, 1996). Hence, we normalized the citation counts based on the following definition and used it for the analysis at the aggregate level:

$$ncit_i = \frac{Total\ citation\ count\ in\ year\ i}{(2010 - year\ i + 1) * number\ of\ papers\ in\ year\ i}$$

where (2010 – year i + 1) represents the gap between the current year and the final year of the study and is used for normalizing the citation counts. The reason for normalizing the number of citations is that older articles have more chance to be cited. Hence, in general as we move forward toward the recent periods the total number of citations decreases. We also used the average impact factor of the journals in which the articles were published as another proxy for the quality of the papers and defined the fourth dependent variable (*avgif*). The last dependent variable represents number of authors per article in year i ($aut\_per\_art_i$) as a measure for the team size of the researchers.

The independent variables that were considered in all the aforementioned models are as follows:

- Small World (*sw*)
- Network Connectivity (*netcon*)[23]

In order to calculate the small world variable, we needed to calculate clustering coefficient and average path length. In the following, the definitions of the clustering coefficient and path length along with the independent variables' definitions are presented.

Clustering Coefficient (CC): This index counts the number of triangles in the 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 in a graph are connected to each other; hence it measures the tendency of the nodes to cluster together (Hanneman & Riddle, 2011). According to Watts and Strogatz (1998) the clustering coefficient can be defined based on a Local Clustering Coefficient (LCC) for each node within a network. LCC is defined as follows:

[23] Control variable.

$$LCC_i = \frac{number\ of\ triangles\ connected\ to\ node\ i}{number\ of\ triples\ centered\ on\ node\ i}$$

The denominator of the above formula counts the number of sets of two edges that are connected to the node *i*. The overall clustering coefficient is calculated by taking average of the local clustering coefficient of all the nodes within the network. Hence,

$$CC = {}^{1}/_{n} \sum_{i=1}^{n} LCC_i$$

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.

Shortest Path Length (PL): This index represents the separation degree of the network and is the lowest number of vertices that are needed to be traversed to reach from one vertex to another vertex (De Nooy, *et al.*, 2005). The shorter the distance is the more easily information may flow among the researchers. The path length was calculated for the largest component of each of 12 created co-authorship networks. From the definition, the small world variable is measured for the largest component of each network. This limitation is due to the fact that the shortest path can be calculated just in a connected network). Hence, we considered the largest connected component[24] for measuring the aforesaid variable in each of the 12 generated networks. This assumption has been widely employed in the literature (*e.g.* Fleming, *et al.*, 2007; Uzzi & Spiro, 2005; He, *et al.*, 2009; Newman, 2000; Liu, *et al.*, 2005; Baum, *et al.*, 2003) and is justifiable, since the core research activities mainly occur in the largest component in which the most influential authors are present (Fatt, *et al*., 2010). Moreover, the proportions of the largest component in our created networks are not only large in comparison with similar studies (*e.g.* Kumar & Jan, 2013; Yan, *et al.*, 2010; Nascimento, *et al.*, 2003; Liu, *et al.*, 2005), but they are even gradually increasing. After 2002 our largest component covered more than 75% of the whole network, reaching to the level of almost 90% in the last period (Figure 43). We can therefore use the largest component for the calculation of the path length.

[24] Component of a network is a sub-network in which there is no isolated vertex and all the vertices are interconnected.

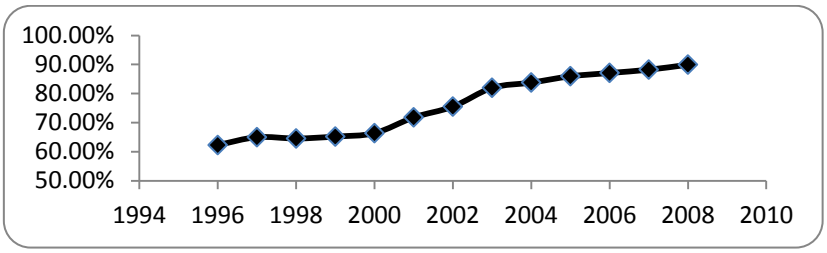


Figure 43. Historical trend of largest component proportion

Small World (SW): The small world variable is calculated based on the clustering coefficient and the path length:

$$SW = {}^{CC}/_{PL}$$

Network Connectivity (netcon): It is a measure of the connections between pairs of vertices and is related to the average degree of the network. In other words, in the co-authorship network of the researchers it indicates the average number of collaborators for each researcher who had at least one article co-authorship during the given period of time. This is an important measure since higher number of co-authors in a network results in a tighter network that facilitates the knowledge exchange (Wasserman, 1994). We used the network connectivity (*netcon*) as our control variable. The reason is that higher number of researchers in a network can increase the chance of higher network connectivity and consequently the chance of higher collaboration among the researchers that may have an effect on our dependent variables. In the following section the results of the study are presented.

5.2.2.3 Results

5.2.2.3.1 Pre-Analysis

Number of the researchers in each of the examined periods of time reflects the size of the network in the corresponding year. As the first step, we analyzed the network size and its trend. According to Figure 44, the network size did not change much until 2000 since when it has been steadily increasing with an almost constant positive slope. Since an annual increase was expected in the number of researchers, the steady line indicating the number of researchers between 1996 and 2000 might be due to the Scopus data that seems to be more integrated and complete for the recent years.

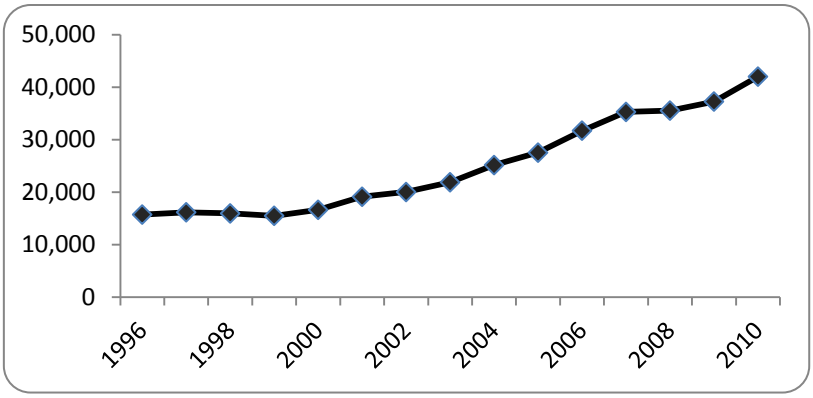


Figure 44. Historical trend of the researchers from 1996 to 2010

In line with the increase in the number of authors an increase is seen in the number of articles, having almost the same trend. According to Figure 45, the number of articles remained constant during the first and the last 5-year periods. However, a positive jump is observed during the second 5-year period (from 2001 to 2005).

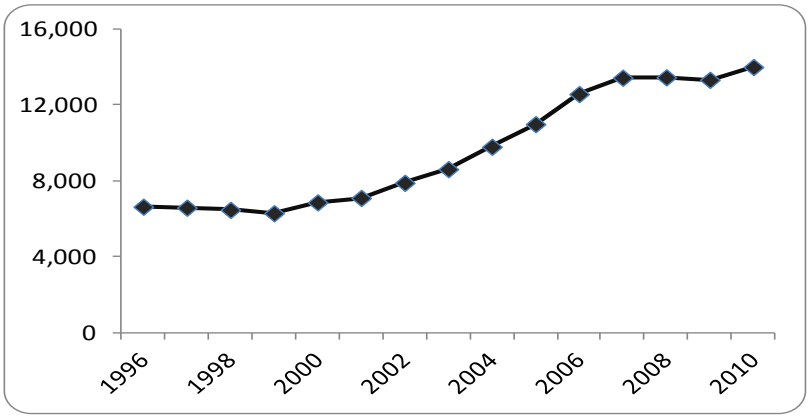


Figure 45. Historical trend of the researchers' articles from 1996 to 2010

5.2.2.3.2 Small World Analysis

According to Kogut and Walker (2001), a network has a small world structure if its average clustering coefficient is significantly higher than a random network of the same number of vertices while having approximately the same path length. Hence, in order to investigate the small world structure in the co-authorship network of the researchers, we constructed an Erdős–Rényi random network of the same size as the actual network for each of the examined periods. The respective path lengths and clustering coefficients were then calculated for the generated random networks and compared to the corresponding amounts of the actual networks. The results are depicted in Figure 46 and Figure 47.

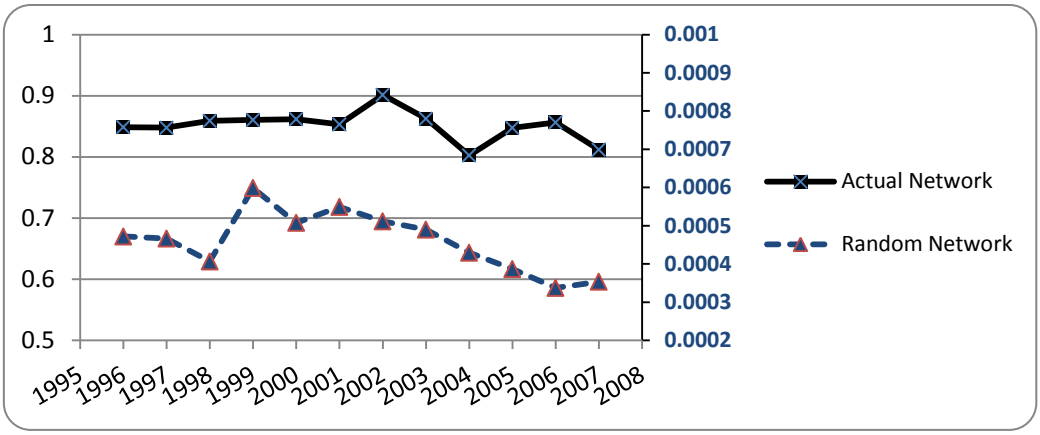


Figure 46. Clustering coefficient, actual and random networks[25]

Although the small world networks are often large in size, they exhibit relatively short path length and high clustering coefficient (Albert & Barabási, 2002). Clustering coefficient in co-authorship network represents the willingness of a researcher's collaborators to collaborate with each other in form of writing a paper jointly (Barabâsi, *et al*., 2002). As it can be seen in Figure 46, the clustering coefficient for the actual network is almost constant maintaining about 0.8 and is significantly higher than the clustering coefficient for the respective random networks (that are between 0.0003 and 0.0006) in all the examined periods. This result is completely in line with the previously done studies that investigated the small world structure (*e.g.* Barabâsi, *et al.*, 2002; Yan, *et al.*, 2010). This is a primary sign of the small world structure in the examined network of researchers. In addition, the clustering coefficient of the examined network is very high in comparison with the other similar studies, *e.g.* all the four co-authorship networks studied by Newman (2001c)[26], and SIGMOD co-authorship networks of Nascimento *et al.* (2003)[27]. This indicates that in the examined network it is more likely for two co-authors to have a common collaborator with whom they have also published an article.

[25] The X-axis in Figure 46 and Figure 47 represents the starting year of each of the 3-year time intervals that were considered to calculate the collaboration network variables. For example, 1996 represents the period of [1996-1998].
[26] The largest clustering coefficient obtained was 0.726.
[27] The largest clustering coefficient obtained was 0.69.

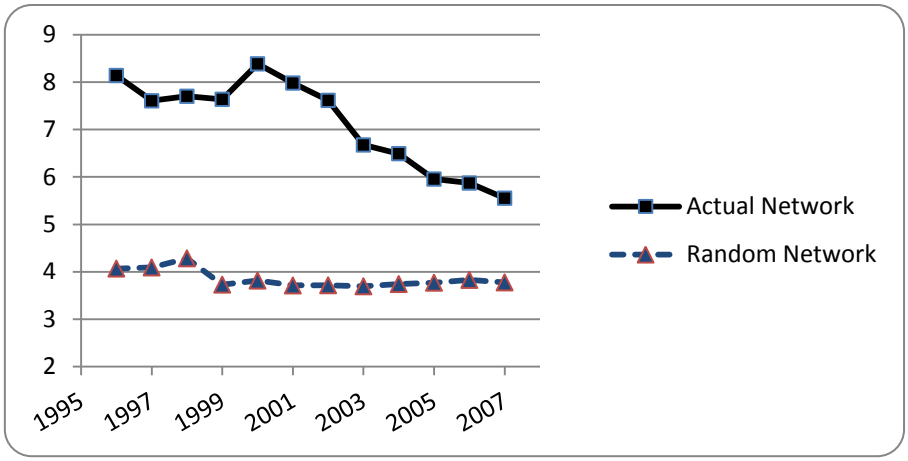


Figure 47. Path length, actual and random networks

We compared the path length for the actual and generated random networks. According to Figure 47, although the path length of the examined co-authorship network remains relatively constant during the initial 5-year period, it starts dropping significantly and continuously after 2000, while getting very close to the path length of the random network. The value of the path length of our examined network is almost similar to the one of Nascimento *et al.* (2003) who found a path length of 5.65 in the SIGMOD co-authorship network, and is lower than some other studies (*e.g.* Liu, *et al.*, 2005). In general, in other similar studies that contain more than 10,000 vertices and analyzed the small world property in co-authorship networks, the average path length is not more than 10 (*e.g.* Newman, 2001a; Newman, 2001b). According to Figure 46 and Figure 47 and based on the definition of Watts and Strogatz (1998) the examined co-authorship network of researchers strictly resembles the small world structure.

As the next step, SW indicator was defined and used to analyze the small world characteristics of the collaboration network of researchers. To calculate the value of the small world indicator we followed the method employed in several similar studies (*e.g.* Davis, *et al.*, 2003; Kogut & Walker, 2001; Baum, *et al.*, 2003) that used the following formula for calculating the small world ratio:

$$SW = \frac{\frac{CC_a}{CC_r}}{\frac{PL_a}{PL_r}}$$

Table 15. Small world characteristics for the collaboration network

		Actual to Random Ratio		
Period	**Network Size**	**Path Length**	**Clustering Coefficient**	**SW**
[1996-1998]	32,862	2.00	1798.74	899.12
[1997-1999]	33,111	1.86	1817.13	977.91
[1998-2000]	33,931	1.80	2113.11	1,175.71
[1999-2001]	36,700	2.05	1436.80	701.86
[2000-2002]	39,870	2.20	1697.40	772.46
[2001-2003]	43,348	2.15	1553.13	722.69
[2002-2004]	47,793	2.05	1762.30	860.31
[2003-2005]	53,191	1.81	1760.39	974.64
[2004-2006]	59,427	1.73	1868.85	1,077.50
[2005-2007]	65,344	1.58	2192.22	1,388.19
[2006-2008]	69,868	1.53	2538.09	1,655.32
[2007-2009]	73,518	1.47	2295.90	1,562.00

Table 15 shows the results for the small world variables calculated for all the examined periods. According to Baum *et al.* (2003), as the size of the network increases the value of the small world indicator should increase. As it can be seen in Table 15, there is an increase in the amount of SW indicator during the first three periods. After a sudden drop, it continues to increase steadily after 1999 reaching to the maximum value of the SW indicator in the latest periods. The drop could be due to two reasons. First, Scopus data was probably less complete during the first intervals, the number of articles found in Scopus is almost constant in the first three periods. Second reason could be the nature of the collaboration network that may have been less mature during the initial periods. As more researchers join the network, more links are established and the network evolves dynamically. This enables the network to reflect more small world properties as the time passes. This proposition is also supported by the trend of the clustering coefficient.

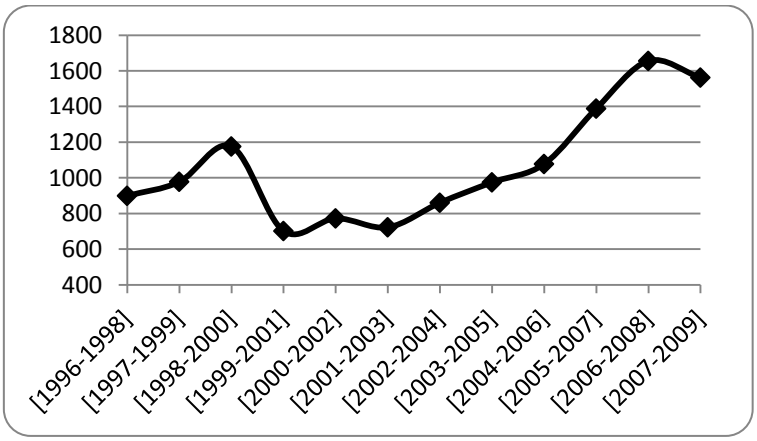


Figure 48. Small world trend

It is also argued in the literature that small world properties follow the form of an inverted U-shape (*e.g.* Gulati, *et al.*, 2012). That means an increase in the small world properties will be followed by a later decrease. According to Figure 48, the trend of SW indicator in the examined network had a local maximum in the period of [1998-2000] and then after a sudden decline it started to rise again till the period of 2006-2008 where the second local maximum is seen. Hence, a declining trend is expected to be seen after 2007 and a reassessment of the small world properties is suggested for the future. In a small world network, researchers can get access to the pools of knowledge in diverse clusters and communities through knowledge brokers who are the actors in the network that connect different clusters. Therefore, other actors can retain or even improve their position in the network by accessing continuously to the flows of diverse information and knowledge or other resources (Eisenhardt & Tabrizi, 1995; Lin, 2002). The reason for the inverted U-shaped form of the small world property is that as the network evolves the knowledge brokers become less important gradually due to the limited advantages of the brokerage positions that will lead to the decline of the small world. In other words, as the network evolves different clusters gradually get familiar with the information pools of the other clusters through the existing knowledge brokers, hence making the knowledge generated in different clusters more homogeneous. Facilitating the knowledge exchange reduces the diversity in the whole network gradually (Lazer & Friedman, 2007), making the role of the knowledge brokers less important. As a result of the decline in the entrance of new knowledge brokers along with the decay of the old brokers, network becomes more separated. Hence, actors prefer to collaborate with their stable and familiar partners within their own clusters and communities. This will lead to multiple isolated clusters and consequently lower small-worldiness (Gulati, *et al.*, 2012).

To compare the small world structure in the examined collaboration network a list of previously identified small world co-authorship networks as well as the network properties is presented in Table 16. Considering the network size, the NSERC researchers' co-authorship network is similar to the SPIRES and LANL co-authorship networks of Newman (2001c), and MATH co-authorship network of Barabási *et al.* (2002). NSERC network is significantly more cliquish (*i.e.* it has a very high respective clustering coefficient) than SPIRES network where the value is quite comparable to the one for the LANL network. However, it is less

cliquish than the MATH network that has the highest clustering coefficient among all the listed networks. Comparing the path length of our examined network with the mentioned networks, it can be said that NSERC network is more similar to the LANL network of Newman (2001c). As it can be seen in Table 16, all the previously studied small world networks have the path length ratio lower than 2, ideally closer to 1. In the case of our examined co-authorship network the path length ratio is declining and getting very close to 1 in the final period (1.47). However, different clustering coefficient ratios are observed in the previous studies that led them to a wide range of values for the SW indicator.

Table 16. Comparison of previously studied co-authorship networks with the last period of our co-authorship network (NSERC)

		Actual to Random Ratio			
Network	**Network Size**	**Path Length**	**Clustering Coefficient**	**SW**	**Reference**
SIGMOD co-authorship	1,413	1.33	172.5	129.7	Nascimento *et al.* (2003)
NCSTRL co-authorship	11,994	1.16	1653.34	1425.3	Newman (2001c)
LANL co-authorship	52,909	1.23	2388.9	1942.2	Newman (2001c)
SPIRES co-authorship	56,627	1.89	242	128.05	Newman (2001c)
Math co-authorship	70,975	1.16	10925.93	9418.91	Barabási *et al.* (2002)
Sociologists co-authorship	128,151	1.30	0.94	0.72	Moody (2004)
MEDLINE co-authorship	1,520,251	0.94	6000	6382.98	Newman (2001c)
NSERC co-authorship	73,518	1.47	2295.90	1,562.0	

5.2.2.3.3 Regression Analysis

After observing the small world structure in the examined co-authorship network, we statistically analyzed the effect of the small world property on several network performance measures. As the first step, we checked for any pair wise correlations among the independent variables and found no significant correlation among them. We considered negative binomial regression model for our first dependent variable, *i.e.* number of articles in the following year. Since the dependent variable in the first model is a count measure, the best regression model would be the Poisson model (Hausman, *et al.*, 1984). However, for a Poisson regression we would need the variance and mean of the sample not to differ significantly. Hence, the data should be tested to detect any over-dispersion or under-dispersion that will lead the Poisson model to underestimate or overestimate the standard errors resulting in misleading estimates for the statistical significance of variables (Coleman & Lazarsfeld, 1981). Therefore, we did the likelihood ratio test to see if the Poisson model fits our data. The results show that the over-dispersion coefficient (α) is significantly

different from zero, which means that Poisson distribution is not an appropriate choice and negative binomial regression could be a better estimator. For the remaining 4 dependent variables, *i.e.* normalized citation count in the following year, average impact factor of journals in which the articles have been published in the following year, number of articles per author in the following year, and number of authors per article in the subsequent year, we used linear regression models.

Table 17 shows the results for the impact of the small world property on productivity of the researchers in terms of the number of their publications. The results show that both of the independent variables (small world and network connectivity) are significant predictors of the scientific productivity in the following year.

Table 17. Regression results for number of articles model

Negative binomial regression — Number of obs = 12; LR chi2(2) = 38.39; Prob > chi2 = 0.0000; Pseudo R2 = 0.1708

Dispersion = mean; Log likelihood = -93.176699

no_art	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
SW	.0003522	.0000558	6.32	0.000	.0002429	.0004615
netcon	.0269058	.0019361	13.90	0.000	.0231111	.0307005
_cons	7.658743	.0934756	81.93	0.000	7.475534	7.841951
/lnalpha	-5.757512	.4220285			-6.584673	-4.930351
alpha	.003159	.0013332			.0013814	.007224

Likelihood-ratio test of alpha=0: chibar2(01) = 327.04 Prob>=chibar2 = 0.000

According to the results, the small world property and network connectivity have a positive impact on the number of publications of the researchers in the subsequent year. These results were expected since as the network becomes more connected, researchers get more familiar with other scientists' fields of research that may lead to the establishment of more collaboration links. In addition, the small world structure can accelerate the exchange of knowledge and expertise among the researchers that may result in higher productivity. The reason is that small world networks allow access to distant information and the knowledge is transferred more efficiently in such networks (Uzzi, *et al.*, 2007). Our results are in accordance with major conclusions of the previous studies (*e.g.* Kogut & Walker, 2001; Cowan & Jonard, 2004; Eslami, *et al.*, 2013).

We checked if the small world network encouraged collaboration among the researchers. We focused on the number of authors per articles as a proxy of the researchers' team size and assessed the effect of the small structure on it. As it can be seen in Table 18, only the small world variable is significant reflecting a small positive impact on the team size. Hence, it seems that researchers benefit from the shorter path length and more clustered sub-networks to get in touch with other researchers who are working in the same scientific area. This may result in the establishment of new collaboration links and expand their team size. Moreover, high clustering creates more repeated links among the researchers, causing the risk to be shared among the researchers that might lead to an increase of the trust level in the community (Chen & Guan, 2010). As the next step, we assessed the impact on the average productivity of the researchers.

Table 18. Linear regression results for team size model

Source	SS	df	MS	Number of obs =	12
				F(2, 9) =	9.74
Model	.194938377	2	.097469189	Prob > F =	0.0056
Residual	.090073661	9	.010008185	R-squared =	0.6840
				Adj R-squared =	0.6137
Total	.285012038	11	.025910185	Root MSE =	.10004

aut_per_art	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SW	.0003812	.0000967	3.94	0.003	.0001624	.0006
netcon	.0034309	.0034135	1.01	0.341	-.004291	.0111527
_cons	2.050345	.1626485	12.61	0.000	1.682408	2.418281

Since the number of authors has an increasing trend over the examined period, to assess the productivity of the researchers more accurately we examined the average number of articles per author. The result of the linear regression model is depicted in Table 19. As it can be seen the small world property has a negative effect on the average productivity of the researchers, which is an interesting finding. Although the small world structure had a positive impact on the total number of articles, it harms the average publication rate. Hence, it seems that in a small world structure researchers start to collaborate more by forming bigger scientific teams that may lead them to increased overall productivity. However, when it comes to the average productivity per researcher it becomes lower since the team sizes have grown. The other aspect to be analyzed is the quality of the papers that are produced. Therefore, in the rest of the section we analyze the impact of the small world network on the quality of the papers.

Table 19. Regression results for average number of articles per author model

Source	SS	df	MS		
Model	.003751416	2	.001875708	Number of obs =	12
Residual	.001715226	9	.000190581	F(2, 9) =	9.84
				Prob > F =	0.0054
Total	.005466641	11	.000496967	R-squared =	0.6862
				Adj R-squared =	0.6165
				Root MSE =	.01381

art_per_aut	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SW	-.0000525	.0000133	-3.93	0.003	-.0000827	-.0000223
netcon	-.0005074	.000471	-1.08	0.309	-.001573	.0005582
_cons	.4630127	.0224446	20.63	0.000	.4122395	.5137859

Two linear regression models were considered to check the impact of small world on the quality of the publications, one is based on number of citations the articles received, and one based on average impact factor of the journals in which the articles were published. Table 20 shows the regression results for the impact of small world structure on the normalized number of citations received in the subsequent year. We have normalized the number of citations based on the year of publication since generally older articles have higher total number of citations.

Table 20. Linear regression results for number of citations model

Source	SS	df	MS		
Model	1.93930954	2	.969654768	Number of obs =	12
Residual	.582490083	9	.06472112	F(2, 9) =	14.98
				Prob > F =	0.0014
Total	2.52179962	11	.229254511	R-squared =	0.7690
				Adj R-squared =	0.7177
				Root MSE =	.2544

ncit	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SW	.0006591	.000246	2.68	0.025	.0001027	.0012155
netcon	.0348163	.0086805	4.01	0.003	.0151796	.0544529
_cons	.8278689	.4136143	2.00	0.076	-.1077916	1.763529

According to the results, the linear regression is well fitted to our data. In addition, both variables are significant at the level of 95% confidence and based on the resulting R^2 the independent variables are relatively good predictors of the dependent variable. Controlling for the network connectivity, small world property has a positive impact on the quality of the papers in the following year in terms of the number of citations received. Hence, it can be said that researchers benefit from the small world structure to exchange ideas more easily, and since they get connected to other researchers they can improve the quality of their work

by internal referring among the team members and other researchers in the network. This is consistent with other studies that analyzed the impact of network centrality measures (not specifically small world properties) on the quality of the papers measured by number of citations and found positive relations (*e.g.* Yan, *et al.*, 2010).

We performed the same analysis using a different proxy for the quality of the papers, namely the average impact factor of the journals in which the articles were published. According to Table 21, a significant positive relation is observed between the average journal impact factor and the small world structure. This along with our findings from Table 20 confirms the importance of the small world structure in producing higher quality publications. From the results it can be said that although small world network may harm the average rate of publications, it will increase the overall quality of the teams' publications.

Table 21. Linear regression results for impact factor model

Source	SS	df	MS		Number of obs =	12
					F(2, 9) =	14.27
Model	19.6831928	2	9.8415964		Prob > F =	0.0016
Residual	6.20607683	9	.689564092		R-squared =	0.7603
					Adj R-squared =	0.7070
Total	25.8892696	11	2.35356997		Root MSE =	.8304

avgjif	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SW	.0037786	.0008029	4.71	0.001	.0019623	.0055948
netcon	.0383485	.0283341	1.35	0.209	-.0257476	.1024446
_cons	2.807872	1.350081	2.08	0.067	-.2462237	5.861967

5.2.2.4 Conclusion

This study focused on the co-authorship network of the Canadian researchers in engineering and natural sciences and investigated the existence of the small world structure and its impact on their productivity, quality of publications, and team size. Several previous studies analyzed different co-authorship networks and found correlations between network centrality measures and researchers' productivity (*e.g.* Yan & Ding, 2009; Abbasi, *et al.*, 2011; Kumar & Jan, 2014; Eslami *et al.*, 2013), however to our knowledge no study has focused specifically on the impact on small world properties on the quality of the publications and scientific team size.

Our results show that the examined network exhibits significant small world properties by having very high clustering coefficient in comparison with the random networks of the equal size while the path lengths are almost the same. The separation degree among scientists decreases to around five in the final period, when it becomes even lower than famous Milgram's (1967) finding of six degrees of separation. Hence, the networks in the final periods become more connected and the low path length among the researchers enables them to exchange knowledge more easily. Moreover, in comparison with most of the other co-authorship networks that have been studied, our examined co-authorship network has relatively larger clustering coefficient, smaller average path length, and larger proportion of the largest component. Specifically, the size of the largest component is critical since the path length (and consequently the small world measure) can be only calculated in the connected sub-network. Hence, this study benefited from the large share of its largest component to have better estimations of the small world variables. On the other hand, the enormous largest component in our examined co-authorship network may represent the fact that the core research activity is being done in an inter-connected large cluster of the researchers. Of course, the size of the largest component also depends on the nature of the research activity and the level of its interdisciplinarity.

According to the results, although the small world structure has a positive impact on the total number of publications, it influences negatively the average productivity of the researchers. Since a positive impact of the small world on the researchers' team size was observed, it can be concluded that researchers may benefit from the small world properties to get familiar with other active researchers in their field and expand their scientific team. This team expansion can bring them several advantages such as internal referring, better and faster access to expertise and other resources, new sources of funding, *etc*. that will result in higher rate of publication for the whole team. But since the size of the team has grown up the average productivity will become lower. Therefore, it seems that even though a small world network would not cause an increase in the individual productivity of the researchers, they will invest their efforts in a more efficient way. Being involved in larger teams and getting in contact with other experts in the field allow them to not only gain new skills but also employ their skills more efficiently. Tighter collaboration among the team members can also create a synergy among them that will surely result in higher productivity of the team. The positive

impact of the small world on the papers' quality also supports the idea that the small world structure facilitates more effective exchange of knowledge among the team members that may result in higher quality work. However, as discussed before, small world properties were reported to follow the form of an inverted U-shape (Gulati, *et al.*, 2012). According to our results, the Canadian natural science and engineering network has seen its latest pick in the period of 2006-2008, after which the article production started to decrease. Hence, according to Gulati *et al*. (2012) a decreasing trend is predicted for the years after 2010, resembling an inverted U-shape curve.

5.2.2.5 Limitations and Future Work

The main limitation was in regard with the sample size. The reason that we selected the time interval of 1996 to 2010 was that Scopus has a weaker coverage before 1996. Moreover, articles need at least three years to be well cited and as a result we did not include the periods after 2010. Future work can address this limitation by using other databases. More observations would allow analyzing the interrelations between the small world property and other network centrality measures to assess the combined impacts.

Another limitation was in regard with the calculation of the small world variable for which we considered the largest component. Although there are some suggestions in the literature for overcoming this limitation (*e.g.* Schilling & Phelps, 2007), they could be applicable when special purpose customizable software for social network analysis is available to code a program to calculate the small world indicator over the whole network. However, as mentioned before the proportion of the largest component in this study was larger than other similar studies, which allowed us to make more realistic estimates of the small world measures.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among the researchers as we were unable to capture other links that might exist among the researchers like informal relationships. These types of connections are never recorded and thus cannot be quantified, but there are certainly some knowledge exchanges occurring in such associations that could affect the network performance. In addition, there are also some drawbacks in using co-authorship as an indicator of collaboration since collaboration does

not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997). Hence, future work can address this issue by taking other indicators into the account.

5.2.3 How the Influencing Factors Affect Researchers' Collaborative Behavior?

Scientific collaboration is one of the substantial drivers of research progress that may lead researchers to generate novel ideas. Scientists may present such new thoughts in high quality journal publications or in the form of technology advances. There are several studies that examined collaboration networks or impact of network variables on scientific activities. However, to our knowledge this paper is the first that analyzes the impact of other influencing factors on network structure variables at the individual level. For this purpose, we focus on the collaboration network among the NSERC funded researchers during the period of 1996 to 2010 and employ time related statistical models to estimate the impact on the 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 local influencers who possess high closeness centrality are not necessarily prolific researchers in terms of the quality of their publications. However, productive researchers who publish high quality articles have higher betweenness and eigenvector centralities. Moreover, although mid-career scientists have higher cliquishness and closeness centrality, the role of young gatekeepers is confirmed in connecting different communities and information spread.

5.2.3.1 Introduction

Recent progress in information technologies has cut the distance world-wide 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 that encourages scientists to be more collaborative in an aim to increase their productivity. However, collaboration may not necessarily augment the scientific performance and several issues need to be considered, *e.g.* selecting the right partner, coordination costs.

Katz and Martin (1997) define scientific collaboration as the process through which the researchers with a common goal work together to produce new scientific knowledge. Scientific collaboration has been studied in a vast number of different disciplines such as computer science, sociology, research policy, and philosophy (Sonnenwald, 2007). Moreover, due to the various types of collaboration mentioned in the literature, *e.g.* inter-firm collaboration, international collaboration, and academic collaboration (Subramanyam, 1983), a diversity is observed in employing various methods, approaches and terminologies in examining the scientific collaboration (Sonnenwald, 2007). Through collaboration researchers get access to an often informal network of scientists that may facilitate knowledge and skill diffusion (Tijssen, *et al.*, 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).

The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006), where financial investment can change the structure of research groups and affect the collaboration among the scientists. However, there might be some conflicts between individual preferences and the society level goals. These conflicts may cause different optimal individual collaboration level from the optimal social one. As a result, to evaluate the policies that affect the collaboration the relation between the individual incentives and social benefits should be considered as well (Ubfal & Maffioli, 2011). Although governmental funding for knowledge creation and diffusion has a long history, its effects on scientific collaboration and formation of scientific networks is relatively new (Katz & Martin, 1997; Lee & Bozeman, 2005).

Researchers have started evaluating the impact of funding on the collaboration using simple indicators in the early 80s (*e.g.* Beaver & Rosen, 1979; Heffner, 1981). Recently, this area attracted the attention of researchers again. Using econometric techniques and statistical analyses in some cases, a few studies assessed the impact of funding and other influencing factors (*e.g.* gender, past productivity, *etc.*) on collaboration during the past ten years. Although some studies found a positive relation between funding and the scientific collaboration (*e.g.* Bozeman & Corley, 2004; Adams, *et al.*, 2005; Gulbrandsen & Smeby,

2005; Defazio, *et al.*, 2009), there also exits few studies that could not find any significant relation between funding and collaboration (*e.g.* Rosenweig, *et al.*, 2008).

This study extends the literature in two ways. To our knowledge, no study has examined the impact of a group of influencing factors on the individual indicators of the position of researchers within their scientific collaboration network. In addition, most of the studies used a limited dataset and/or focused on a limited scope while this study uses a large dataset of the funded researchers. Our basic motivating questions are: How the influencing factors including funding affect the position of the scientists among their collaboration network? What are the most determinant factors in stimulating scientific collaboration? The remainder of the paper proceeds as follows: Section 5.2.3.2 presents the data, methodology and the model; Section 5.2.3.3 presents the empirical results and interpretations; Section 5.2.3.4 concludes; and Section 5.2.3.5 discusses the limitations.

5.2.3.2 Data and Methodology

5.2.3.2.1 Data

The data of this research was gathered in three phases. In the first phase, the funded researchers' data was extracted from NSERC and then using Elsevier's Scopus we collected all the information (*e.g.* co-authors, their affiliations, year of publication) about the articles that were published by the funded researchers within the period of 1996 to 2010. The main reasons for selecting NSERC was its 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), and we decided to focus on the period of 1996 to 2010 since the data quality of Scopus was lower before 1996. In addition, 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. SCImago was chosen since it provides annual data of the journal impact factors that enables us to perform a more accurate analysis as we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. In addition, SCImago is powered by Scopus that makes it more compatible with our publications database.

In the second phase, we did a full text search over the articles and fetch the ones that acknowledged NSERC support in the body of the articles. 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. This will surely result in an over-estimation of the number of articles. The procedure that we took is based on the assumption that all the grantees should acknowledge the source of funding in the article. The refined data from phases one and two was integrated into a single MySQL table.

In the last phase of the data gathering procedure, we used Pajek software to construct the co-authorship networks of the funded researchers for each single year of the examined time interval and to calculate the network structure variables at the individual level. The calculated network structure indicators were integrated into the database. The final database contains 174,773 records. In the next section, we discuss the methodologies used in this research.

5.2.3.2.2 Methodology

As discussed in the previous section, we first employed social network analysis to construct the collaboration network of the funded researchers and to measure the structural network properties. As the next step, we used statistical analysis to analyze the impact. For this purpose, we considered four different dependent variables that are average team size of the funded researchers (*teamSize*) measured by the average number of authors per paper, betweenness centrality (*bc*), clustering coefficient (*cc*), eigenvector centrality (*ec*), and closeness centrality (*cl*). Number of authors per paper has been used in the literature as a proxy for scientific collaboration (*e.g.* Beaver & Rosen, 1979; Rosenweig, *et al.*, 2008). The definition of the other dependent variables is presented in the rest of this section.

Betweenness Centrality (*bc*) focuses on the role of intermediary individuals in a network. The 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 two other nodes in a network, the higher betweenness centrality it has that 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:

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

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 the 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 connected to each other (Hanneman & Riddle, 2011). Watts and Strogatz (1998) defined clustering coefficient based on a Local Clustering Coefficient (*lcc*) for each node within a network. The definition of *lcc* is:

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

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

$$cc = {}^{1}/_{n} \sum_{i=1}^{n} LCC_i$$

in which *n* denotes 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 the network depends also on the importance of its connections. Hence, an actor is more central if it is connected with other actors who are themselves central. In other words, eigenvector centrality measures how well connected an actor is in the network. Bonacich (1972) defined the centrality of an actor based on the 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 hence, it is expected that they shape the collaborations and play an important role in setting priorities on scientific projects.

Closeness Centrality (*cl*) was first proposed by Sabidussi in 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. Hence, the closeness centrality of a node *i* in a graph with *N* nodes is:

$$cl_i = \frac{1}{\sum_{j \in N-\{i\}} d(i,j)}$$

where *d(i,j)* is the length of the shortest path between the nodes *i* and *j*. Based on the definition, closeness centrality can only be calculated in connected components (graphs) since if the graph is not connected the denominator becomes infinity and as a result the closeness centrality would be zero which is not informative. We calculated this centrality measure in the largest connected component of the co-authorship networks.

To perform the statistical analysis, a regression model was defined for each of the dependent variables and STATA 12 data analysis and statistical software was used to estimate the models. The reduced form of the regression models is as follows:

$$\begin{bmatrix} teamSize_i \\ bc_i \\ cc_i \\ ec_i \\ cl_i \end{bmatrix} = f(avgFund3_{i-1}, avgIf3_{i-1}, avgArt3_{i-1}, avgCit3_{i-1}, dc_i, careerAge_i, d_i)$$

In the regression models, $avgFund3_{i-1}$ is the average amount of funding that a researcher has received over the past three years. In the literature three-year (*e.g.* Payne & Siow, 2003) or five year (*e.g.* Jacob & Lefgren, 2007) time windows have been considered for the funding to take effect. We considered both for our model and found that the three-year time window is better suited. 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 that 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 citation count of the articles in the past three years as another

measure for the quality of the papers. Past productivity of the funded researcher is represented by $noArt_{i-1}$ in the model and was measured as the average number of articles for a researcher in a three year time window. 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 equipments, *etc.* may cause the higher productivity. 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 author's first article in the database and the given year.

Degree Centrality (*dc*) variable is also included in the regression models in which the network variables are dependent. This measure is defined based on the number of ties that a node has (degree) in an undirected graph. Hence, degree central researchers (actors) should be more active since they have higher number of ties (links) to other researchers (Wasserman, 1994). Degree centrality of node *i* is defined based on the node's degree and then the values are normalized between 0 and 1 to be able to compare the centralities:

$$dc_i = \frac{degree\ of\ node\ i}{highest\ degree\ in\ the\ network}$$

In each of the models we used different types of dummy variables. The dummy variable $dInst_i$ represents the type of the affiliation of the funded researcher, whether it is affiliated with academia or non-academia environments. For the Canadian provinces, we defined another dummy variable $dProvince_i$. To compare the impact of different NSERC funding programs another dummy variable was defined ($dProg_i$).

5.2.3.3 Results

5.2.3.3.1 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). Figure 49 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 (vertical line in Figure 49) the slope has become steeper indicating a considerable increase in the average

amount of funding. In addition, during the first five years of the examined time interval (dashed vertical line in Figure 49) we see a steadier trend of the average funding in comparison with the other periods. We will use the vertical lines of the average funding in the rest of the figures of this section in order 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.

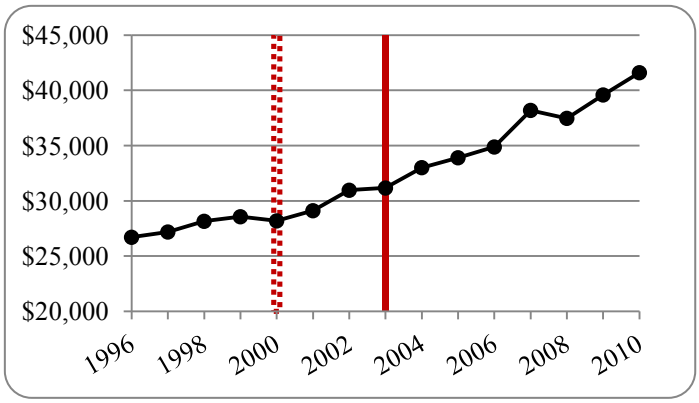


Figure 49. Average funding per researcher

Researchers publish their results in books or journal articles or present them in scientific conferences 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 big science that requires collaboration and expertise of many individuals (De Solla Price, *et al.*, 1986). Number of authors per paper has been considered as a proxy for scientific collaboration in several studies (*e.g.* Newman, 2004; Rosenweig, *et al.*, 2008). Figure 50 presents the average amount of authors per paper for the funded researchers. The vertical lines show different periods of average funding that was discussed earlier. According to Figure 50, as the amount of the average funding increases the average number of authors per articles also augments. In other words, it seems that higher funding enables funded researchers to form larger scientific teams in an aim 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).

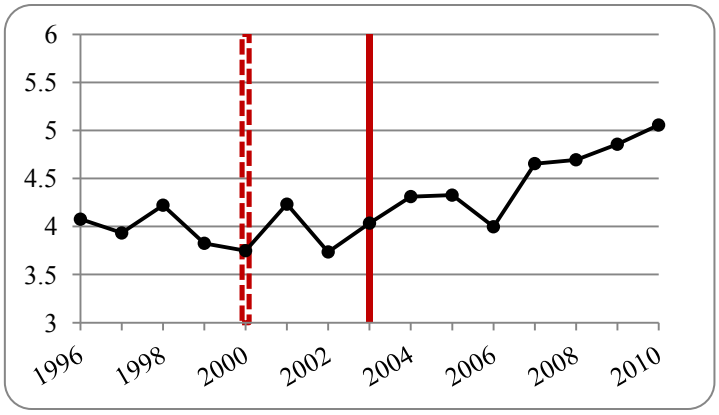


Figure 50. Average number of authors per article

Trends of the network structure variables are represented in Figure 51. 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 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 Figures 49, 50, and 51 it seems that at the aggregate level there is a positive relation between funding and collaboration measured by the average number of authors per article. However, nothing can be said for the network structure variables. Hence, in order to assess the effects more accurately we turn to the regression analysis and investigate the impact of the influencing factors on collaboration at the individual level.

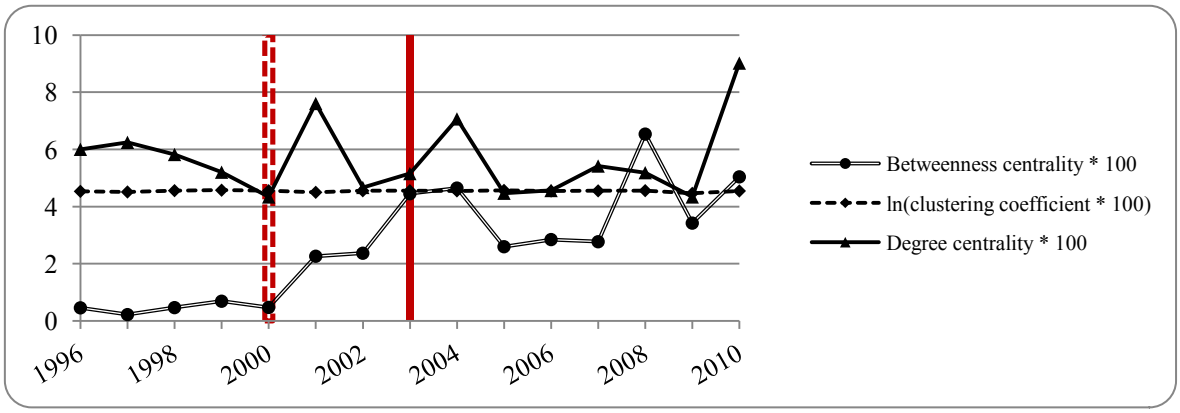


Figure 51. Average betweenness centrality, clustering coefficient, and degree centrality

5.2.3.3.2 Statistical Analysis

As discussed in section 5.2.3.2.2, 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 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 B.

5.2.3.3.2.1 Average number of authors per paper (teamSize)

The impact of the influencing factors on the scientific team size of the researchers measured by average number of authors per article was analyzed at the individual level. We calculated the scientific team size in two ways, one considering distinct co-authors of a researcher (*distinct team size* model) and the other one by taking all the co-authors into the account (*overall team size* model). 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). We used non-linear time related multiple regressions for the analysis purpose. The regression result for the overall team size model is presented in Table 22.

Table 22. Regression result, overall team size model

$teamSize_i$	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf. Interval]*	
$ln\_avgFund3_{i-1}$	1.092452***	.2116682	5.16	0.000	.6775817	1.507322
$noArt3_{i-1}$	.6196625***	.1215581	5.10	0.000	.3814082	.8579168
$ln\_avgCit3_{i-1}$	1.189371***	.1944058	6.12	0.000	.8083347	1.570407
ln_avgIf3_{i-1}	4.353832***	.3167796	13.74	0.000	3.732943	4.974721
$careerAge_i$	-.7124596***	.2184799	-3.26	0.001	-1.140681	-.2842384
$careerAge_i^2$	.0346901***	.013419	2.59	0.010	.0083889	.0609913
Affiliations dummy variable						
dAcademia	-8.096576***	1.12634	-7.19	0.000	-10.30421	-5.888946
_cons	.8180767	2.312413	0.35	0.724	-3.71426	5.350413

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 60,907

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 the 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 affect the scientific collaboration positively.

As expected, the past productivity of the funded researchers measured by the average 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 benefit from larger scientific teams. According to the results not only the rate of publications affects the team size, the quality of the works also positively influences the collaboration (*avgCit3* and *avgIf3*). In other words, higher quality papers of the funded researchers in the past three-year has a positive relation with their scientific team size in the following year. Hence, the results suggest that productive researchers who produce high quality works are more collaborative.

We controlled for the age of the researchers in the regression model and as expected 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, *et al.*, 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 grow 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. 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 (apex at the bottom).

To assess the impact of the type of the affiliation of the researcher on collaboration, the institution type dummy variable (*dAcademia*) was also added to the model that takes value 1 if the funded researcher belongs to the academia environment and 0 if the affiliation is non-academia. As it can be seen, academia funded researchers are significantly different from the non-academia ones and they work in smaller scientific teams in comparison with their non-academic counterparts. We also considered dummy variables for different Canadian provinces and funding programs. Analysis of the provinces dummy variables reveals that the Canadian provinces do not have significantly different impact on the team size of the researchers since none of the provinces was significant. We omitted the discovery grants program and analyzed the impact of funding program dummy variable. According to the

results, just *tools* and *industrial* funding programs are significantly different from the omitted program where both of them had a positive coefficient.

As the next step, we focused on the distinct average team size of the funded researchers and did the same analysis. To calculate the distinct team size we just counted distinct co-authors (collaborators) of a funded researcher. The results are presented in Appendix B. According to the results the sign of the influencing factors are the same as the ones of Table 22 but the coefficients are smaller indicating a lower intensity of the considered factors. Hence, in general the discussion presented for the overall team size model is also valid for the distinct size model.

5.2.3.3.2.2 Network Structure Variables

In this section, the impact of influencing factors on the network structure variables is assessed. For this purpose, four regression models are estimated in which betweenness centrality, clustering coefficient, closeness centrality, and eigenvector centrality are considered as the dependent variables separately. The multiple regression analysis is done at the individual level of the researchers. For all the models, we considered all the combinations of the variables (*i.e.* independent, interaction terms, dummy variables, and quadratic forms of the variables whenever it was meaningful) as well as the lags for the considered variables and present the most robust results. As a proxy of the scientific team size of the researchers, we added the independent variable of degree centrality of the researchers (*dc*). Table 23 shows the regression results for the betweenness centrality model.

Table 23. Regression result, betweenness centrality (bc) model

$bc_i * 10^4$	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
$ln\_avgFund3_{i-1}$	.4350983***	.0653106	6.66	0.000	.307088	.5631086
$noArt3_{i-1}$	1.409226***	.0332145	42.43	0.000	1.344124	1.474327
$ln\_avgCit3_{i-1}$	.9382845***	.0609348	15.40	0.000	.8188508	1.057718
ln_avgIf3_{i-1}	-.2877661***	.1026326	-2.80	0.005	-.4889287	-.0866036
$dc_i * 10^4$	.0097551***	.00219	4.45	0.000	.0054626	.0140476
$careerAge_i$	-.0328585**	.0165532	-1.99	0.047	-.0653031	-.0004139
Affiliations dummy variable						
dAcademia	-.0432513	.3988558	-0.11	0.914	-.8250186	.738516
_cons	-5.563956***	.734255	-7.58	0.000	-7.003114	-4.124798

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 38,974

According to the results, 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 more 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 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 positive impact of the team size of the researchers measured by their degree centrality (*dc*) since higher amounts of funding may enable researchers to expand their scientific activities that might be resulted 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 (*avgFund3*) 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 and according to the results not necessarily publishing in higher quality journals may lead the researcher to a more influential position. As it can be seen in Table 23, 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.

We also compared betweenness centralities of the researchers affiliated with academia and non-academia, estimated by the *dAcademia* dummy variable in the model. According to the results, the affiliation of the 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. We omitted Ontario province and defined dummy variables for the remained nine provinces and found that none of the dummy variables of the provinces became significant at the level of 90%. This confirms that locating in one of the other nine provinces does not have a significant different impact from locating in Ontario on the betweenness centrality of the researchers. Finally, we defined dummy

variables for the most frequent NSERC funding programs, namely discovery grants, strategic projects, industrial funding, collaborative grant, and tools and equipment grants. The dummy variable of the discovery grants was omitted. It was found that the collaborative grants and strategic projects are significantly and positively different from the omitted dummy variable at the level of 90% and 99% respectively. This partially indicates that the researchers who have been funded through collaborative or strategic programs possess in general more central positions in comparison with their counterparts who have been supported by the discovery grants. This finding is completely in line with the definition of the mentioned funding programs. Specifically for the strategic project grants, the aim is to improve the scientific development in selected high-priority areas that influences Canada's economic and societal position. Hence, these well-defined targeted grants should be allocated to specific reputable researchers, probably with more central positions and higher influential potency.

The next network structure variable that we focused on as a dependent variable was the clustering coefficient (*cc*) of the researchers at the individual level (individual cliquishness). 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[28] in total. 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 les than two neighbors. Table 24 shows the results of the non-linear regression analysis for the clustering coefficient. According to the results, funding has a negative impact on the clustering coefficient of the researchers. It can be said that researchers may use the allocated funding to find more new partners rather than forming 3-loops (triangles) among their previous partners. Hence, it seems that more funding will result in linear expansion of the team of the researchers. In addition, it is observed that past productivity measured by the number of articles in the past 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).

[28] A clique is a subset of the vertices of a graph in a way that every two vertices are connected by an edge.

Table 24. Regression result, clustering coefficient (cc) model

cc_i	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf. Interval]*	
$ln\_avgFund3_{i-1}$	-.0072308***	.0023715	-3.05	0.002	-.011879	-.0025826
$noArt3_{i-1}$	-.0131229***	.0003736	-35.13	0.000	-.0138552	-.0123907
$ln\_avgCit3_{i-1}$	.0289128***	.0020735	13.94	0.000	.0248486	.032977
ln_avgIf3_{i-1}	.0130777***	.0034918	3.75	0.000	.0062336	.0199217
$dc_i * 10^4$	.0011079***	.0001433	7.73	0.000	.000827	.0013889
$careerAge_i$	-.0119676***	.0022802	-5.25	0.000	-.0164369	-.0074982
$careerAge^2_i$	.0007043***	.0001377	5.11	0.000	.0004344	.0009742
Interaction variables						
(dc_i*10^4) * $careerAge_i$	-.0000404**	.0000172	-2.34	0.019	-.0000742	-6.60e-06
Affiliations dummy variable						
dAcademia	-.1292246***	.0140584	-9.19	0.000	-.1567794	-.1016698
Provinces dummy variables						
dQuebec	.0354775***	.0051104	6.94	0.000	.0254609	.0454941
dBColumbia	.0055326	.006027	0.92	0.359	-.0062804	.0173456
dAlberta	-.0166545***	.0063172	-2.64	0.008	-.0290364	-.0042727
dSaskatchewan	.014967	.0108709	1.38	0.169	-.0063403	.0362744
dNBrunswick	-.0217324	.0140335	-1.55	0.121	-.0492385	.0057737
dManitoba	.0055321	.0113353	0.49	0.626	-.0166852	.0277495
dNFoundland	-.0014286	.0151796	-0.09	0.925	-.031181	.0283238
dPEdward	.0466338	.0352544	1.32	0.186	-.0224657	.1157332
dNScotia	.0176086*	.0098687	1.78	0.074	-.0017344	.0369516
Funding programs dummy variables						
dStrategic	.030983***	.00817	3.79	0.000	.0149696	.0469963
dTools	.0406788***	.0114563	3.55	0.000	.0182243	.0631334
dCollaborative	.0379478***	.008083	4.69	0.000	.0221049	.0537907
dIndustrial	-.0103649	.0130488	-0.79	0.427	-.0359408	.015211
_cons	.8134445***	.0263064	30.92	0.000	.7618833	.8650057

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 38,974

Both of the proxies for the quality of the papers (*avgCit3* and *avgIf3*) have a positive impact on the clustering coefficient of the researchers. This is also quite reasonable since 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 led 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 the overall impact of the career age becomes positive and 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 that 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 * careerAge*), it can be said that there is a balance between the number of direct partners of a researchers and his/her age. In other words, although it was found mid-career scientists are on average more cliquish, if they have too many direct partners it may affect their cliquishness negatively.

In order to see where the clustering coefficient is higher and to be able to compare the cliquishness, we included dummy variables in the regression model representing for the institution type, Canadian provinces, and different NSERC funding programs. The institution type dummy variable (*dAcademia*) takes value 1 if the researcher belongs to the academia environment and 0 if his affiliation is non-academia. According to Table 24, academia funded researchers are significantly different from the non-academia ones and have on average around 13% (-0.129) less cliquishness in comparison with the non-academic researchers. For the analysis of the Canadian provinces dummy variables, we omitted Ontario. 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 that may indicate higher clustering coefficient of the researchers located in the mentioned provinces in comparison with Ontario.

For comparing the impact of different NSERC funding programs, we categorized the programs into 5 categories: discovery grants, strategic projects, collaborative grants, tools and equipment grants, and industrial funding programs. We decided to omit the discovery grants since it is the most frequent and common funding program among the Canadian researchers. According to Table 24, the effects of strategic, tools, and collaborative funding

programs are significantly and positively different from the discovery grants program. According to the definition of these grants the results are quite reasonable and expected. Specifically for the strategic project grants which has 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 influences Canada's economic and societal position. Hence, this finding also confirms that 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.

The next network variable that we focused on as the dependent variable was eigenvector centrality (*ec*). As explained earlier we tested several models where the most robust results are presented in Table 25. Surprisingly, funding (*avgFund3*) has a negative impact on eigenvector centrality. Hence, it seems 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*. 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 have 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 team size of a scientist, has also a positive effect on the 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 possess high betweenness centrality (acting as gatekeepers) or even high closeness centrality (acting as local influencers). They are highly connected actors with mainly high profile individuals within highly interconnected clusters. Interestingly, the interaction of degree and career age of the researchers represents 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.

Table 25. Regression result, eigenvector centrality (ec) model

$ec_i * 10^4$	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
$ln\_avgFund3_{i-1}$	-.396984***	.0790308	-5.02	0.000	-.5518864	-.2420816
$noArt3_{i-1}$	.048448***	.0133974	3.62	0.000	.0221888	.0747072
$ln\_avgCit3_{i-1}$	.1861356**	.0737369	2.52	0.012	.0416094	.3306618
ln_avgIf3_{i-1}	.0607812	.1242528	0.49	0.625	-.1827574	.3043197
$dc_i * 10^4$	.3065048***	.0051434	59.59	0.000	.2964236	.316586
$careerAge_i$	.0187393	.0201824	0.93	0.353	-.0208187	.0582974
Interaction variables						
$(dc_i*10^4) * careerAge_i$	-.0019642***	.0006198	-3.17	0.002	-.003179	-.0007494
Affiliations dummy variable						
dAcademia	1.183675**	.4826933	2.45	0.014	.237584	2.129766
_cons	1.930254**	.8890298	2.17	0.030	.1877336	3.672775

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 38,974

The analysis of the institution type dummy variable (*dAcademia*) reveals that researchers who work in the academia environment are significantly different from their industrial counterparts. The positive coefficient of the dummy variable indicates that academia researchers are more likely to have higher eigenvector centrality (to act as the leaders) in the co-authorship networks rather than the non-academic scientists.

Finally, we assessed the impact of the influencing factors on the closeness centrality (*cl*) of the researchers at the individual level. For this purpose, we first calculated the closeness centrality of the researchers in the largest component of the co-authorship networks since closeness centrality can be only calculated in connected networks. According to Table 26, average funding (*avgFund3*) positively affects the 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 positive affect 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 that much clear since the citation based proxy (*avgCit3*) shows a negative impact while the journal impact factor based measure (*avgIf3*) presents a positive effect. Hence, it seems that local influencers are not necessarily highly prolific scientists in terms of the quality of their publications. As it was expected, the direct scientific team size of the researchers, measured by degree centrality (*dc*), has a significant 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 the closeness centrality of the researchers is at first negative. 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.

Table 26. Regression result, closeness centrality (cl) model

$cl * 10^2_i$	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf. Interval]*	
$ln\_avgFund3_{i-1}$	.1694727***	.0269733	6.28	0.000	.1166018	.2223436
$noArt3_{i-1}$	.0215427***	.0034334	6.27	0.000	.0148128	.0282727
$ln\_avgCit3_{i-1}$	-.0734994***	.0263145	-2.79	0.005	-.1250789	-.0219198
ln_avgIf3_{i-1}	.4752498***	.0437594	10.86	0.000	.389476	.5610235
$dc_i * 10^2$	2.593725***	.0546285	47.48	0.000	2.486647	2.700804
$careerAge_i$	-.4191376***	.0261731	-16.01	0.000	-.4704401	-.3678351
$careerAge^2_i$	.0243901***	.0015257	15.99	0.000	.0213995	.0273807
Affiliations dummy variable						
dAcademia	.0596839	.1436248	0.42	0.678	-.2218382	.341206
Provinces dummy variables						
dQuebec	.1445022**	.0587166	2.46	0.014	.0294105	.2595938
dBColumbia	.1965733***	.0691152	2.84	0.004	.061099	.3320475
dAlberta	.2391774***	.0722264	3.31	0.001	.0976048	.3807499
dSaskatchewan	.2614194**	.1213772	2.15	0.031	.0235052	.4993335
dNBrunswick	.1175653	.1700892	0.69	0.489	-.2158303	.4509608
dManitoba	.2754394**	.1375401	2.00	0.045	.005844	.5450347
dNFoundland	.1694044	.1903767	0.89	0.374	-.2037573	.542566
dPEdward	-.1685734	.5198238	-0.32	0.746	-1.187491	.8503446
dNScotia	.1549521	.1162304	1.33	0.183	-.0728737	.3827779
Funding programs dummy variables						
dStrategic	.0758724	.0853524	0.89	0.374	-.0914287	.2431735
dTools	.7589592***	.1259065	6.03	0.000	.5121672	1.005751
dCollaborative	.0675027	.0863846	0.78	0.435	-.1018216	.2368271
dIndustrial	-.2931606**	.134483	-2.18	0.029	-.5567636	-.0295576
_cons	8.016092***	.2915756	27.49	0.000	7.444568	8.587615

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 15,046

Dummy variables of three different types were also added to the regression model to assess and compare the effect of institution type of researchers, their residence, and different

NSERC funding programs. As it can be seen academia 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. To compare the impact of residency we defined dummy variables for different Canadian provinces and omitted Ontario. According to the results, 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 reside in Manitoba. We also compared the impact of different funding programs. Same as the analysis that was explained in the previous sections, we omitted the discovery grants program since it is the most frequent and common funding program among the Canadian researchers. As it can be seen in Table 26, the effect is only different for tools and industrial funding programs, with positive and negative coefficients respectively.

5.2.3.4 Conclusion

In this paper we investigated the impact of funding and other influencing factors like past productivity, team size, 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 the dependent variables and performs the impact analysis on 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 productive researchers who are producing high quality papers on average have larger scientific teams. This partially confirms the importance of the collaboration in scientific

activities in a way that productive researchers who tend to produce high quality works also tend to be more collaborative. Analyzing the career age of the researchers showed that the career age of the researchers negatively influences their collaboration that might be due to the difficulties in managing the costs of collaboration (*e.g.* finding right partners and research coordination).

In the second part of the analysis the impact was investigated on the network structure variables. 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 past productivity of the researchers in terms of both quantity and quality of the publications along with the average amount of funding available to them are the most crucial factors in achieving higher betweenness centrality. Analyzing the impact of degree centrality as a measure of team size on betweenness centrality revealed that in the examined co-authorship network higher number of direct connections empowers the role of gatekeepers. Surprisingly, a negative impact was observed for the career age of the researchers on their betweenness centrality that might indicate the considerable role of the 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 that might indicate the linear use of funding resources by researchers in the examined network to expand their direct partners rather than empowering their internal teams through forming triangles. Interestingly, a negative impact was observed for the rate of publication on the cliquishness 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 more triangles (loops) 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 (through, *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 the researchers' first publication, the overall impact of the career age on the clustering coefficient becomes positive. Hence, 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. However, although mid-career scientists are on average more cliquish, the number of their direct connections will be affected by their career age in a way that if they continue to increase their direct partners their cliquishness will be negatively affected after several years.

Analyzing 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 the reputable highly central researchers. Therefore, they can play an important role in forming scientific collaboration teams or in defining new projects and setting priorities on the projects. Surprisingly, a negative impact was observed for the funding on the eigenvector centrality which 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 showed a positive impact on the researchers' leadership role that is quite expected. One 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. This finding was also confirmed by the positive impact of the degree centrality on the eigenvector centrality since higher number of direct connections increase 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 represents a negative effect on the 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 (*cl*) 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 specifically 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 impact of the career age becomes positive after 18 years hence it seems that mid-career scientists are more likely to have higher influence within their local community.

5.2.3.5 Limitations and Future Work

We were exposed to some limitations in this paper. The source of bibliographic data (Scopus) is English biased (that is an inevitable since other similar sources suffer from the same limitation), hence, non-English articles are underrepresented (Okubo, 1997). 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 have been proposed in the literature for calculating the closeness centrality in disconnected graphs (*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.

5.2.4 How to Get more Funding for Research?

Funding has been viewed in the literature as one of the main determinants of scientific activities. Several studies assessed the impact of funding on scientific activities and performance of the funded researchers. However, to our knowledge this article is the first that measures the effect of several important factors (*e.g.* past productivity of the researchers,

scientific collaboration of the researchers, career age of the scientists, *etc.*) on the amount of funding that is allocated to the researchers at the individual level. For this purpose, a time-related non-linear multiple regression model is estimated. Our results suggest a positive relation between collaboration network structure variables and the amount of funding except for the eigenvector centrality. Moreover, our findings show that researchers who are highly productive in terms of both quantity and quality of their papers receive on average higher amount of funding. In addition, funding allocation seems to be biased towards the senior researchers in a way that as a researcher moves forward in his/her career it becomes more likely to secure higher amount of funding.

5.2.4.1 Introduction

About 100 years ago, the power and wealth of the nations were measured by their amount of natural resources or the industrialization stage. Apart from the human capital that is an essential factor for scientific development and innovation (Griffith, *et al.*, 2004), knowledge has also become a new worthy capital and a basis for competitiveness (Klette & Kortum, 2002). In this respect, it is essential to strive to increase the production of the knowledge, which could be estimated by the research outcomes in terms of publication, scientific applications, and income (Oyo, *et al.*, 2008). Funding has been acknowledged as one of the main drivers of scientific activities (Martin, 2003). It can play a significant role in defining new scientific projects and/or setting priorities on the existing projects.

Investment strategies on research and development (R&D) can affect the performance of the funded researchers and their interactions with other scientists. In addition, funding can influence the size and efficiency of R&D sector and its productivity (Jacob & Lefgren, 2011). Higher scientific performance can be reached by better funding allocation through selecting highly prolific research groups or well-defined projects, supporting novel ideas, and targeting structural changes such as promoting scientific collaboration networks (Braun, 2003). However, different nations follow various research patterns and their institutional and economic structures greatly differ. Hence, the composition of the budget which different countries are allocating to R&D activities varies as well. As a result, various allocation patterns are used world-wide to distribute the research funding among universities and research institutes (Leydesdorff & Wagner, 2009).

Governments put significant efforts on defining a systematic framework for evaluating the performance of researchers in regards to the amount of funding that they have been receiving. In addition, policies on the R&D activities have evolved over the past fifty years (Elzinga & Jamison, 1995; Sanz-Menéndez & Borras, 2001). Beginning with the research promotion through public research centers, motivating and incentive mechanisms were introduced during 1960s and 1970s, first by the *research councils* (Rip, 1994) and later through *strategic R&D programmes* (Irvine & Martin, 1984), in order to further stimulate firms and universities to advance their scientific activities. Nowadays, different countries are experiencing various types of governmental interventions and policies. Due to the limited financial resources and importance of the scientific development, therefore, assessing the effectiveness of the government policies as well as the performance of the funded researchers is becoming more vital.

In Canada the importance of receiving research funding is on the rise especially among the academic researchers (Polster, 2007) that could make the competition for getting more grants even tighter. The Canadian government (like most governments in the Western countries (Geuna, 2001)) has focused on the universities as the key research units of the country over the past 25 years in order to secure the national competitiveness worldwide (Polster, 2007). Therefore, several policies have been set (*e.g.* commercialization of university research, setting research priorities, and promoting targeted areas) in order to encourage the academic researchers and to better establish the key role of the universities (Industry Canada, 2002). Changes in federal funding policies, lack of university operating budgets, higher priority of the selected strategic research projects, and rising research costs have made the research grants more important than ever to the Canadian researchers (Polster, 2007).

A lot of studies have analyzed the impact of funding on the productivity and performance of the funded researchers in terms of quantity and quality of their publications at micro (*e.g.* Arora & Gambardella, 1998; Godin, 2003; Payne & Siow, 2003; Jacob & Lefgren, 2007) or macro level (*e.g.* Leydesdorff & Wagner, 2009; Shapira & Wang, 2010). In addition, there exist some studies that focused on the scientific collaboration among the researchers and assessed the impact of funding on the formation or rate of collaboration (*e.g.*

Heffner, 1981; Adams, *et al.*, 2005; Defazio, *et al.*, 2009). However, according to our knowledge there is no study that investigates the impact of influencing factors on the amount of funding that researchers receive. This study considers funding as a dependent variable and systematically analyzes the impact of some determinant factors (*e.g.* past productivity of the researchers, collaboration network variables, career age of the scientists, *etc.*) on funding. Therefore, this paper extends the literature in two ways. Firstly, to our knowledge, no study has identified and examined the factors which determine the allocated funding to the researchers at the individual level. We will address this gap through employing statistical analysis techniques on an extensive dataset. Secondly, it will identify the profile of the highly funded researchers and will shed a light on how a researcher can obtain more funding. Our basic motivating questions in this research are: What factors are important in getting more funding? What are the influencing factors that affect the amount of funding that a scientist receives? The remainder of the paper proceeds as follows: Section 5.2.4.2 presents the data and methodology; Section 5.2.4.3 presents the empirical results and interpretations; Section 5.2.4.4 concludes; and Section 5.2.4.5 discusses the limitations and suggests directions for the future work.

5.2.4.2 Data and Methodology

5.2.4.2.1 Data

The data for this research was gathered in two phases. In the first phase, the funded researchers' data was extracted from NSERC and then using Elsevier's Scopus we collected all the information (*e.g.* co-authors, their affiliations, year of publication) about the articles that were published by the funded researchers within the period of 1996 to 2010. In addition, 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 in for the period of 1996 to 2012. Selecting the period of 1996 to 2012 for the citations of the papers enabled us to consider the citations for each article in a three-year time window. For example if an article was published in 1996 its citation counts were collected and averaged over the period of 1996 to 1998, and for the articles published in 2010 (the latest year in the publications database) citations were collected from 2010 to 2012. In the second phase of the data gathering procedure, we used Pajek software to construct the co-authorship networks of the

funded researchers and to calculate the network structure variables at the individual level and in a three year time window. The calculated network structure indicators were integrated into the database. The final database contains 174,773 records. In the next section, we discuss the methodologies used in this research.

5.2.4.2.2 Methodology

As the first step of the analysis, we used several indicators to initially assess the impact of various influencing factors and analyze their trends. The results will be presented in section 5.2.4.3.1. After the primary descriptive analysis, we employed social network analysis to construct the co-authorship network of the researchers and to measure the structural network properties. More specifically, we will calculate four network structure variables that are betweenness centrality, degree centrality, eigenvector centrality, and clustering coefficient and will assess their impact on funding. The definition of the mentioned network variables are as follows:

Betweenness centrality (*bc*) focuses on the role of intermediary individuals in a network. The betweenness centrality of a node *k* is measured based on the share of times that a node *i* reaches a node *j* via the shortest path passing from the node *k* (Borgatti, 2005). In our co-authorship network, the more often a researcher lies on the shortest path between two other researchers in the network, the higher betweenness centrality it has. High betweenness centrality of the node thus indicates the high control of that researcher over other two non-adjacent researchers (Wasserman, 1994). Hence, betweenness centrality of node *k* (bc_k) is defined as follows:

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

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*.

Degree centrality (*dc*) is defined based on the number of ties that a node has (*i.e.* degree of the node) in an undirected graph. Researchers who has high degree centrality can be more active since they have higher number of ties to other researchers (Wasserman, 1994). Hence,

degree centrality for node *i* is defined based on the node's degree and then the values are normalized between 0 and 1 to be able to compare centralities:

$$dc_i = \frac{degree\ of\ node\ i}{highest\ degree\ in\ the\ network}$$

Eigenvector centrality (*ec*) is based on the idea that the importance of a node in the network depends also on the importance of its connections. Hence, in our co-authorship network a researcher will have higher eigenvector centrality if he/she is connected with other scientists who are themselves central. In other words, eigenvector centrality measures how well connected a researcher in the network is. According to Bonacich (1972) the centrality of a node is defined based on sum of its adjacent centralities. In our network, we name researchers who have high eigenvector centrality values as *leaders* since they are connected with too many other influential and highly central researchers. We expect that they shape the collaborations among researchers and play an important role in setting priorities on scientific projects. Hence, we will also assess the impact of eigenvector centrality on funding.

The last network structure variable that we evaluate its impact on funding is the clustering coefficient (*cc*). This measure is also called cliquishness in the literature and it is defined as the likelihood that two neighbors of a node are also connected to each other (Hanneman & Riddle, 2011). Watts and Strogatz (1998) defined clustering coefficient based on a local clustering coefficient (*lcc*) for each node within a network. The definition of *lcc* is:

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

The denominator of the above formula counts the number of sets of two edges that are connected to the node *i*. The overall clustering coefficient is calculated by taking average of the local clustering coefficient of all the nodes within the network. Hence,

$$cc = {}^{1}/_{n} \sum_{i=1}^{n} LCC_i$$

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).

Apart from the network structure variables that represent several aspects of scientific collaboration among the researchers, other measures (*e.g.* productivity of the researchers, quality of the papers, *etc.*) were calculated and integrated into the statistical model as independent variables along with the network variables. For each of the independent variables a three-year time window was considered and the impact of them was evaluated on the average amount of funding in the following year at the individual level. As an example, for the funding year of 1999 we construct the co-authorship network of the researchers during the period of 1996 to 1998 and calculate the network structure indicators for the mentioned three-year sub-networks. The three-year time window for calculating the network structure variables has been already used in the literature (*e.g.* Beaudry & Allaoui, 2012). STATA 12 software package was used to perform the statistical analysis. The reduced form of the regression model is as follows:

$$\boldsymbol{fund_i} = \boldsymbol{f(avgIf3_{i-1}, avgCit3_{i-1}, noArt3_{i-1}, bc3_{i-1}, dc3_{i-1}, cc3_{i-1}, ec3_{i-1}, carAge_i, d_i)}$$

We considered two proxies for the quality of the papers, one is based on the citation counts and the other based on the impact factor of the journals in which the articles were published. Both of them can serve as a proxy for quality, but with a slightly different meaning. Impact factor indicates the respectability of the journal, *i.e.* the quality and the level of contribution perceived by the authors and the reviewers of the paper, whereas the citations show the impact of the articles. Both proxies have some flaws, so we decided to include both of them. We added $avgIf3_{i-1}$ to the model that is calculated based on the average impact factor of the journals that 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 number of citations for the articles in the past three years as another measure for the quality of the papers. Past productivity of the funded researcher is represented by $noArt3_{i-1}$ in the model and is measured as the average number of articles for a researcher in a three year time window. Four network structure variables that were defined earlier are calculated in a three-

year time window (*bc3*, *dc3*, *cc3*, *ec3*) and integrated into the model reflecting different characteristics of the scientific collaboration networks.

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 equipments, *etc.* may cause the higher productivity. Hence as a proxy for the career age of the researchers, we included a control variable named $carAge_i$ representing the time difference between the date of the first article of a researcher in the database and the given year. We used different types of dummy variables in our regression model that were represented in general by d_i in the proposed reduced form of the regression model. The included dummy variables are defined based on the type of the affiliation of the researchers, the Canadian provinces, and their involvement in the largest connected co-authorship sub-network.

5.2.4.3 Results

5.2.4.3.1 Descriptive Analysis

Before analyzing the regression model, we first examine the trends of some related indicators to provide a general picture. Funding is regarded as the main determinant influencing factor of scientific activities (Martin, 2003). Figure 52 shows the average amount of NSERC funding granted to distinct individual researchers from 1996 to 2010. As indicated by the red dashed line in the figure the average funding has followed an increasing trend during the examined time interval reaching from the level of $32,000 in the first considered year to around $49,000 in the final period.

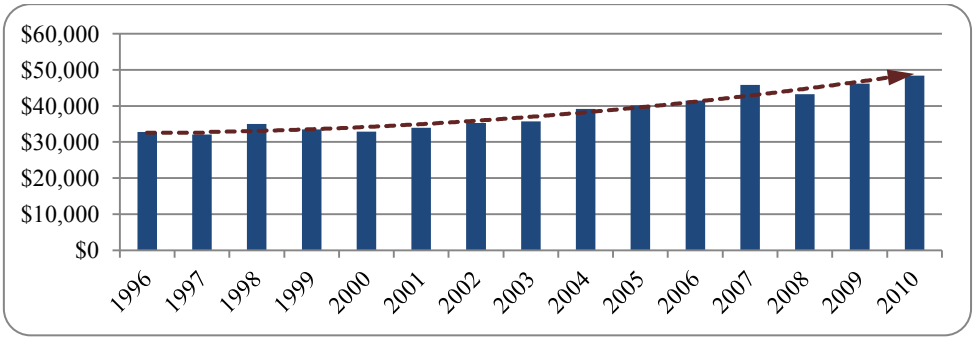


Figure 52. Average funding per distinct researcher, 1996 to 2010

Researchers publish their results in books or journal articles or present them in scientific conferences in order to ensure priority for their discoveries and raise their scientific reputation. Number of publications has been widely used in the literature as a proxy for scientific output. Figure 53-a depicts the average number of papers per researcher normalized between 0 and 1 during the examined time interval. The trend can be divided into two parts as indicated by the vertical dashed line in the figure that are decreasing trend from 1996 to 1999 and increasing trend afterwards. As it can be seen, till 1999 the average number of articles per researcher is declining while after 1999 it starts to increase. The slope becomes steeper after 2003 and it continues till 2007 while after a sudden drop in 2008 it continues to augment with almost similar slope. Figure 53-b shows the overall relation between the amount of average funding and the number of publications. Intuitively it seems that there is a positive relation between funding and scientific output.

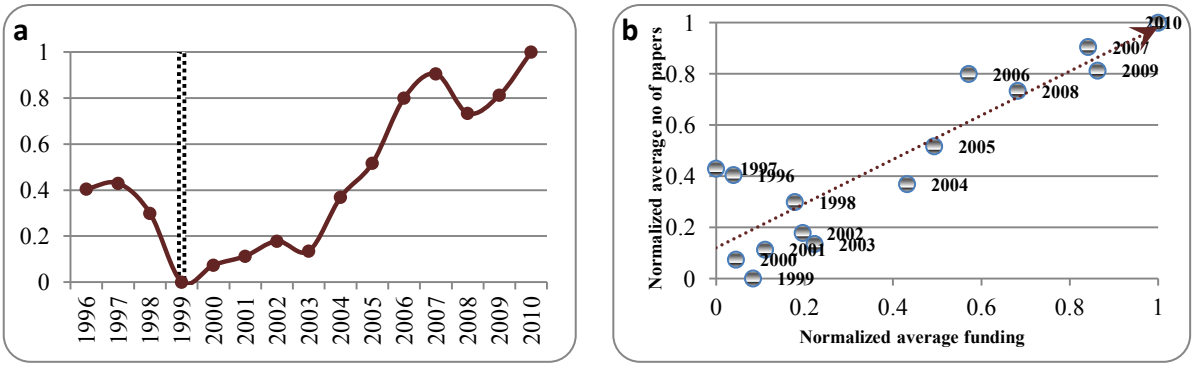


Figure 53. a) Normalized average number of papers per researcher, 1996 to 2010, b) Normalized average number of papers versus normalized average funding, 1996 to 2010

Apart from the rate of publications we have also analyzed the trend of their quality. As mentioned earlier, number of citations received by an article and the impact factor of the journal in which the article was published are the two most common measures for the quality of the paper. However, it is argued that journal impact factor cannot be considered as a good paper quality measure since it is highly discipline dependent and editorial policies can also affect the impact factor (Moed, *et al.*, 1996; Seglen, 1997). Number of citations has also some drawbacks (*e.g.* negative citations and self-citation) but citation based indicators are considered as the common practice in measuring the overall impact of an article (Seglen, 1992). We defined a three year time window for both funding and articles to calculate the average amount of citations. For example as it can be seen in Figure 54, for the funding year

of 1996 we collected all the articles of the funded researchers for the period of 1996 to 1998. Then, we defined a three-year citation window for each of the publication years. In other words, we counted the citations for the period of 1996 to 1998 for the articles that were published in 1996, and from 1997 to 1999 for the articles published in 1997, and from 1998 to 2000 for the articles published in 1998. We followed the same procedure for the other funding years and in order to make a fair indicator we stopped at the funding year of 2008 since we had the publications for the period of 1996 to 2010 and the citations for the period of 1996 to 2012.

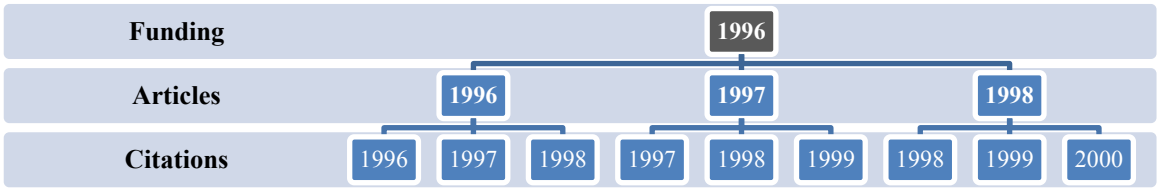


Figure 54. Example of the procedure for counting the citations received by the articles

Figure 55-a depicts the trend of 3-year average citation indicator over the period of 1996 to 2008. As it can be seen, the overall trend follows an increasing polynomial curve of degree 4 (the dashed curve in the figure). As indicated by the dashed vertical lines, the trend can be divided into three regions. Except for the period of 2002 to 2005 for which we see an almost steady trend, in the other parts the average number of citations has increased. The slope is much steeper for the period of 1998 to 2002. Figure 55-b shows the normalized average citations received by the articles versus the average amount of funding allocated to the researchers labeled for different years. As it can be observed, no relation between funding and quality of the papers can be observed in the figure. For example, for the period of 1996 to 2003 that is shaded in Figure 55-b, although the annual average amounts of funding are comparable (see only a very slight increase in Figure 52) a considerable difference is seen in the amount of citations. This is a preliminary result and we will further investigate this issue by the statistical analysis.

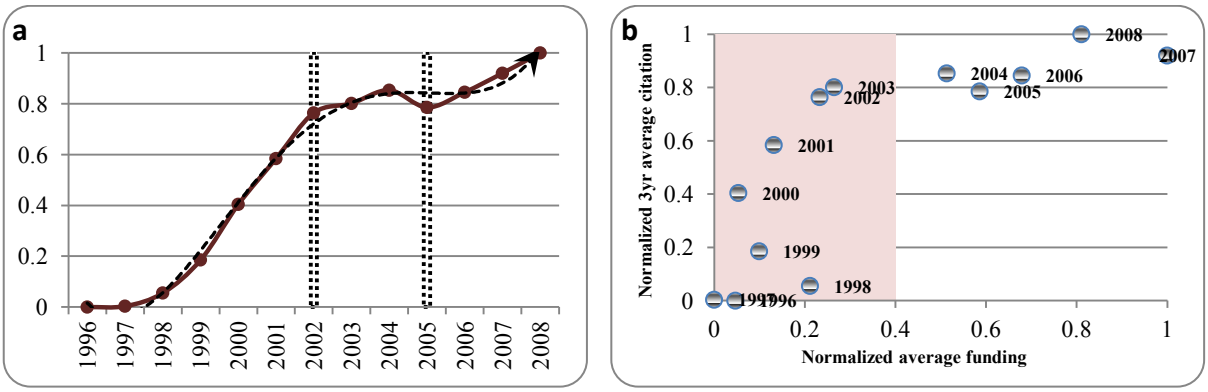


Figure 55. a) Normalized 3-year average citation counts, 1996 to 2008, b) Normalized 3-year average citation counts versus normalized average funding, 1996 to 2008

Analyzing the trend of network structure variables in a three-year time window at the aggregate level reveals that within the period of [2000-2002] till [2001-2003] all the examined network variables were relatively high (the shaded area in Figure 56). Hence, it seems that no relation exists between the network variables at the aggregate level and the amount of average funding as the trend of funding has been slightly increasing during the whole examined period (Figure 52). One reason for the drop in the values of the aggregate network variables in the recent years can be the increasing trend of the involvement of new researches in the network. We will investigate the impact of network variables more accurately by calculating them at the individual level and assessing their effect statistically in section 5.2.4.3.2.

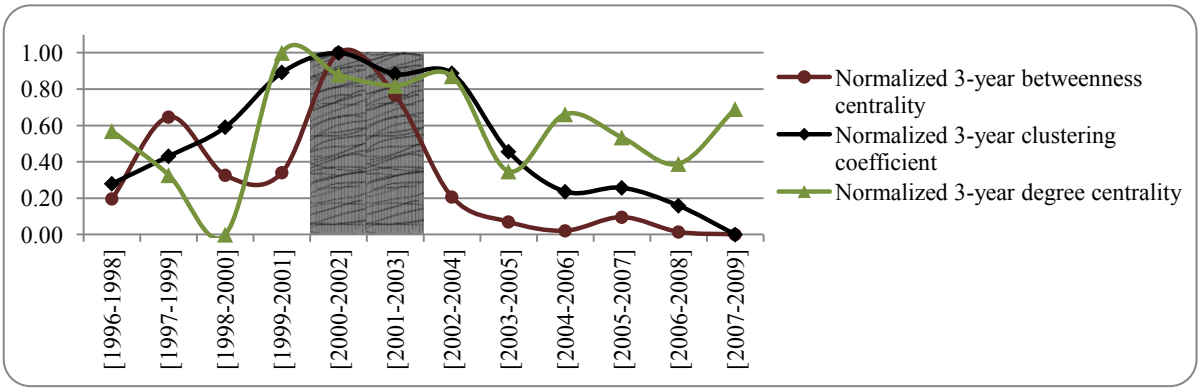


Figure 56. Network structure variables at the aggregate level

We also examined the interaction of the career age of the scientists with the average amount of funding allocated to them. For this purpose, we searched over our publications database from 1991 to 2010 for each of the scientists in the database and set career age of a

researcher to one on the date that he/she produced his/her first publication. Hence, for the period of 1991 to 2010 the career age ranges from 1 to 20. Having set the career age of the researchers, we focused on the range of 1996 to 2010 and compared the amount of funding allocated to the researchers of different career ages. According to Figure 57, it can be said that there is a positive relation between the career age of the researchers and the amount of funding that they have received until the age of 15. However, some fluctuations are observed after the career age passes 15 reaching to a minimum at the age of 20. Hence, it seems that as the researchers start their career the funding allocated to them is minimal at first, but continues to increase and peaks at a certain age of their career. We will assess the impact of the career age of scientists on funding in the statistical analysis more accurately.

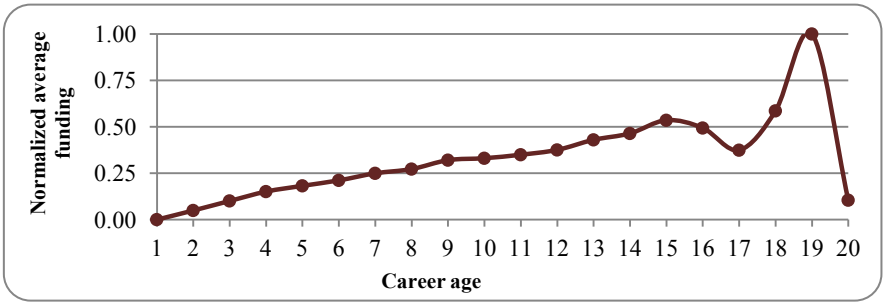


Figure 57. Normalized average funding per distinct researcher versus career age

5.2.4.3.2 Statistical Analysis

In this section we statistically analyze the impact of the proposed influencing factors on the amount of funding allocated to the researchers at the individual level. As explained in section 5.2.4.2.2. four network structure variables (*i.e.* betweenness centrality (*bc*), clustering coefficient (*cc*), eigenvector centrality (*ec*), and degree centrality (*dc*)) along with two measures for the quality of the papers (based on journal impact factor (*avgIf*) and citations counts (*avgCit*)), as well as the number of publications (*noArt*) and the career age of the researchers (*carAge*) were considered as the independent variables. We considered all the researchers who are affiliated with universities, research institutes, and industrial firms and performed multiple regression analysis at the individual level. Moreover, we filtered the data to include only the researchers for whom all the network structure variables could have been calculated. This resulted in 72,267 records of the researchers.

Before running the regression model, we first analyzed the associations between dependent and independent variables. We considered all the combinations of the one-year, two-year, and three-year 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). According to Table 27, the absolute value of all the correlation coefficients is lower than 0.46, which indicates that the degree of linear correlation among the selected variables is weak. For most of the interactions the degree of linear correlation is significantly very weak.

Table 27. Correlation matrix, funding model

Variable	$kFund_i$	$noArt3_{i-1}$	$avgIf3_{i-1}$	$avgCit3_{i-1}$	$bc3*10^2_i$	$dc3*10^2_i$	$cc3_i$	$ec3*10^2_i$	$carAge_i$
$kFund_i$*	1.0000								
$noArt3_{i-1}$	0.3085	1.0000							
$avgIf3_{i-1}$	0.0683	0.0425	1.0000						
$avgCit3_{i-1}$	0.0598	0.0361	0.2894	1.0000					
$bc3*10^2_i$	0.1767	0.4591	0.0396	0.0243	1.0000				
$dc3*10^2_i$	0.0759	0.1365	0.1456	0.1098	0.0880	1.0000			
$cc3_i$	-0.0894	-0.3570	0.0332	0.0454	-0.1760	0.0439	1.0000		
$ec3*10^2_i$	0.0099	0.0423	0.0474	0.0166	-0.0039	0.4248	0.0309	1.0000	
$carAge_i$	0.1598	0.3239	0.0063	0.0544	0.1168	-0.0002	-0.2447	-0.0032	1.0000

* Note: *kFund* is the amount of funding divided by 1000.

As mentioned, we considered a time interval as well as a lag structure of one, two and three years for most of the variables in our model and tested all the combinations of the lags and time intervals and selected the ones that had the most robust results. Hence, three year time interval was found to be the most appropriate for all the independent variables that were tested and one-year lag was found suitable for the past productivity variables. We also added the career age of the researchers to the model as the control variable. In order to see the curvature of its impact the quadratic term was also included in the model. The results of the multiple regression analysis are shown in Table 28. In the rest of this section we take each of the independent variables in turn and evaluate its effect on the amount of funding that researchers receive.

According to the results, the rate and quality of the publications in the past three years have a positive impact on the amount of funding in the following year. Among them the number of publications has the highest impact while the effect of the average number of citations is the lowest. Hence, it can be said that researchers who are highly productive in terms of the quantity and quality of their papers receive on average higher amount of

funding. It was argued in some studies (*e.g.* Zucker, *et al.*, 2007; Beaudry & Allaoui, 2012) that higher amount of funding available will result in higher number of publications. Here we found that the other direction of the equation could be also true.

Table 28. Regression result, funding model

$kFund_i$	*Coef.*	*Std. Err.*	*t*	*P>|t|*	*[95% Conf. Interval]*	
$noArt3_{i-1}$	5.548908***	.0961418	57.72	0.000	5.360471	5.737346
$avgIf3_{i-1}$	3.659511***	.3372827	10.85	0.000	2.998438	4.320584
$avgCit3_{i-1}$	.258806***	.0376524	6.87	0.000	.1850074	.3326046
$bc3*10^2_i$	45.58743***	4.312701	10.57	0.000	37.13455	54.04031
$dc3*10^2_i$	27.94798***	3.750963	7.45	0.000	20.59611	35.29986
$cc3_i$	8.185887***	1.103446	7.42	0.000	6.023136	10.34864
$ec3*10^2_i$	-14.57707***	3.202601	-4.55	0.000	-20.85416	-8.299987
$careerAge_i$	.8142736*	.4190712	1.94	0.052	-.0071046	1.635652
$careerAge^2_i$	.057353**	.0263626	2.18	0.030	.0056824	.1090236
Largest component dummy variable						
dInLargest	15.3695***	.9099608	16.89	0.000	13.58598	17.15302
Affiliations dummy variable						
dAcademia	4.822487**	2.145102	2.25	0.025	.6180933	9.026881
Provinces dummy variables						
dQuebec	6.233359***	.9871349	6.31	0.000	4.298578	8.168141
dBColumbia	9.389953***	1.181226	7.95	0.000	7.074754	11.70515
dAlberta	-1.783301	1.26717	-1.41	0.159	-4.266951	.7003479
dSaskatchewan	-4.504571**	2.081214	-2.16	0.030	-8.583745	-.4253972
dNBrunswick	-6.073225**	2.560385	-2.37	0.018	-11.09157	-1.054878
dManitoba	-10.56926***	2.179204	-4.85	0.000	-14.8405	-6.298031
dNFoundland	-12.21***	2.819985	-4.33	0.000	-17.73717	-6.682842
dPEdward	-26.7532***	6.724452	-3.98	0.000	-39.9331	-13.57329
dNScotia	-5.472374***	1.904909	-2.87	0.004	-9.205989	-1.738759
_cons	15.83564***	2.621302	6.04	0.000	10.6979	20.97339

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 72,267

Interestingly, the impact of average journal factor is higher than the impact of the average number of citations. A possible explanation may be related to the reputation of the researchers possibly affecting their success in publishing in high impact factor journals. We assume that the scientists who are well known and more recognized in their scientific community have on average higher chance of publishing articles in higher quality journals. Therefore, having high average of journal impact factors can also partially reflect the likelihood of being more reputable. This is also partially confirmed by the positive impact of the career age on the amount of funding that will be discussed later. In addition, it is quite in line with Arora and Gambardella (1998) who suggest that well-known highly reputable

researchers with an established record of successes are more likely to receive higher amount of funding.

Network structure variables reflect the impact of collaboration patterns and researchers' position in the co-authorship network on the amount of funding that they receive. According to Table 28, betweenness centrality (*bc*) has a significant positive impact on the amount of funding. A researcher with high betweenness centrality is playing an important role in the network as he/she lies on a relatively high proportion of shortest paths between other researchers. Hence, researchers would have to go through the researcher with high betweenness centrality to reach other researchers. Therefore, these highly central researchers can control the flow of knowledge and can influence the formation and evolution of scientific teams and research projects acting as "*gatekeepers*". Based on these explanations, the positive relation between funding and betweenness centrality was quite expected.

Degree centrality can be regarded as a proxy for the scientific team size of the researchers in the co-authorship network. In other words, a researcher with high degree centrality has on average higher number of co-authors in comparison with the counterparts with lower degree centrality. Therefore, they can have better access to other researchers that might enable them to get involved in more projects. In addition, they have more knowledge sources, so a better access to knowledge in general. This can enable them to come up with a higher variety and more interesting research ideas and project proposals. The quality of proposals is supposed to be one of the main factors for the funding allocation. Moreover, a degree central researcher can also be regarded as a "*social*" researcher who is in contact with a relatively high number of other researchers that might enable him/her to be aware of resource transactions among other researchers, hence increasing the chance of being involved in new projects and/or securing new funding resources. In contrast, a "*peripheral*" researcher has on average few or even no relations, which lowers his/her chances to meet other potential researchers, get involved in high priority well-defined projects, and secure new funding resources. According to Table 28, our results also suggest a positive impact of the degree centrality on the amount of funding that researchers receive.

As it can be seen in Table 28, clustering coefficient (cc) has also a positive impact on funding. As mentioned earlier, clustering coefficient is a measure of the number of triangles

(cliques) in a network, and is also called the *cliquishness*. In the co-authorship network, a researcher with high clustering coefficient has on average a more connected neighborhood in a sense that if his/her neighborhood is fully connected (*i.e.* there exists a connection among all the researchers in the neighborhood) then his/her clustering coefficient would be one. As the number of connections in the neighborhood decreases the value of the cliquishness gets closer to zero. The positive relation between cliquishness and funding shows the importance of being involved in well connected communities. That means apart from the important positive role of the number of direct connections (degree centrality) on funding, being a member of a better connected community also increases the chance of securing more money. A researcher in a more integrated *clique* is more likely to be involved in a more multidisciplinary research that needs interaction among all members of the team. Hence, our results partially suggest that working in a multidisciplinary project can also increase the chance of getting more money for the research. The complex nature of modern science forces researchers to go beyond the restricted circle of their direct connections and get involved in more interdisciplinary research by which they can get access to novel skills and expertise and even new financial resources.

Eigenvector centrality takes the inter-connectivity of a researcher's connections into the account in a way that a researcher who is connected to more central and important researchers obtains high eigenvector centrality. Hence, eigenvector centrality is a more global network analysis measure since it considers the overall structure of the network. Based on our results a negative relation is observed between the eigenvector centrality of the researchers and the amount of funding (Table 28). Observing a negative impact of the eigenvector centrality along with the positive effect of the other examined network structure variables that were already discussed may indicate that in order to secure higher level of funding it would be better for the researchers to try to be directly connected to a lot of people and work in big teams, to be always included in the tight community and assure that the information flows through them, but getting connected to highly influential and important people (leaders) can harm their amount of funding since in this case they may lose the lead and the other important researchers may take it over. Interestingly, the absolute value of the coefficients for all the network variables are much higher than the past productivity

variables. This may indicate more importance of building the collaboration network and informal relations rather than productivity in securing more amount of funding.

Evaluating the effect of the career age of the researchers reveals the positive relation between the age and the amount of funding. In general, as the career age of the researchers grows they gain more reputation in the scientific community. In addition, as they move forward they acquire more experience in writing funding proposals and searching for new funding resources. Moreover, their collaboration network becomes more connected gradually. Hence, older scientists tend to receive higher amount of funding.

Several dummy variables were added to the model to compare the effect of various categorizers on the amount of funding. As it can be seen in Table 28, being in the largest component of the co-authorship network (estimated by the *dInLargest* dummy variable) can be advantageous for the researcher in securing higher amount of funding. It is quite expected since according to the definition the largest component of a network is a connected sub-network with the largest number of vertices. Being a member of the largest component means that the researcher is a part of a connected network hence it would be more likely for him/her to secure more sources of funding as he/she can on average reach more researchers (directly or indirectly) in comparison with an isolated researcher or a researcher in a smaller connected sub-network.

We also evaluated the effect of the institution type of the researchers measured by *dAcademia* dummy variable that takes value 1 if the researcher's affiliation is academic and 0 if his/her affiliation is non-academic. Our results suggest that academic researchers are significantly different from the non-academic ones and are on average more likely to receive higher amount of funding. We checked for the impact of being located in different Canadian provinces by including provinces dummy variables in the model. For this purpose we omitted Ontario and defined dummy variables for the remaining nine Canadian provinces to compare their impact with Ontario. All the provinces dummy variables were significant at the level of 95% except for Alberta. Interestingly, researchers who are located in Quebec and British Columbia tend to receive more amount of funding in comparison with the researchers of Ontario. However, the other researchers who are located in Saskatchewan, New Brunswick, Manitoba, Newfoundland and Labrador, Prince Edward, and Nova Scotia

provinces are on average receiving lower amount of funding in comparison with their counterparts who are located in Ontario. This partially highlights different importance and priority of the Canadian provinces and their universities and research institutes in regard to the amount of funding that NSERC is allocating to their researchers.

5.2.4.4 Conclusion

This paper analyzed the impact of various influential factors (*e.g.* past productivity of the researcher, collaboration network, career age, *etc.*) on the amount of funding that researchers receive. We employed social network analysis and non-linear multiple regression model to assess the impact of the considered factors on funding at the individual level. A three-year time window was considered for the past productivity (in terms of both quantity and quality of the publications) and network structure variables to show their impact on the amount of funding in the following year. To our knowledge this is the first comprehensive study that considers the network structure variables along with several other factors and evaluates their impact on researchers' funding at the individual level.

As discussed, the number of direct co-authors of a researcher may reflect the extent of the opportunities that is available to him/her to collaborate such as getting involved in new projects, exchanging knowledge with other skillful scientists, getting access to new funding resources, *etc*. Hence, as it was expected a positive relation was seen between the degree centrality and funding. On the other hand, occupying more central positions in terms of betweenness centrality and cliquishness can be also beneficial for the researcher. Having higher betweenness centrality can bring a strategic importance to the researcher that might result in higher amount of funding. Higher cliquishness can provide a highly connected local network for the researcher that might open the gate to new financial resources especially in multidisciplinary scientific fields. Hence, researchers' effort in forming more cliquish local collaboration networks will be rewarded by extra amount of funding. However, as it was observed higher eigenvector centrality has a negative impact on funding that indicates the lower importance of having indirect highly influential collaborators in securing higher level of funding.

Analyzing the effect of past productivity of the researchers on funding revealed a positive relation between both quantity and quality of their papers on the amount of funding that they receive. Hence, it can be said that more productive researchers are more likely to receive higher amount of funding. In addition, it was observed that as the career age of the researchers grow the amount of grants also increases. Therefore, it is more probable for the senior researchers to secure higher amount of funding in comparison with their junior counterparts. This finding was quite expected since as the career age of the researchers grows they get on average more reputation in the scientific community that they work while their collaboration network also becomes more established. Moreover, senior researchers might be more experienced in writing funding proposals and applying for new grants.

Apart from the important role of the network structure variables and researchers' position in the collaboration network, this paper highlighted the significant role of being connected to other researchers in securing higher amount of funding as the researchers who are in the largest connected component of the co-authorship network receive on average higher amount of funding than isolated scientists or the ones who are in smaller sub-networks. In addition, it was observed that academic researchers are more likely to receive higher amount of funding rather than the researchers who are affiliated with non-academic environment. Finally, according to the results Canadian provinces can be divided into two groups namely, high and low funding provinces. Ontario, Quebec, British Columbia, and Alberta can be assigned to the high funding group of provinces where the researchers who are located in the mentioned group receive on average higher amount of funding. Within the high funding provinces, it was observed that researchers from Quebec and British Columbia receive on average higher amount of funding than their counterparts in Ontario while no significant difference was observed for the amount of funding between the researchers in Ontario and Alberta. The other six Canadian provinces belong to the low funding group of provinces. Researchers located in the low funding group of provinces receive on average lower amount of funding from all the provinces in the high funding group.

Due to the limited resources of funding and the increasing number of researchers, the competition among researchers to secure the required funding for their research is becoming tighter. This study can be considered as a guideline for the researchers who are seeking to

secure more amount of funding for their research projects. In addition, it was observed that grants and funding allocation is more biased towards senior researchers. Hence, it would be suggested to set new strategies in favor of young productive researchers.

5.2.4.5 Limitations and Future Work

We were exposed to some limitations in this paper. Scopus was selected 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). In addition, since Scopus data were 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.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among the researchers as we were unable to capture other links that might exist among the researchers like informal relationships. These types of connections are never recorded and thus cannot be quantified, but there are certainly some knowledge exchanges occurring during such associations that could affect the network performance. In addition, there are also some drawbacks in using co-authorship as an indicator of scientific collaboration since collaboration does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997). Hence, future work can address this issue by taking other types of collaboration networks into the consideration.

5.3 Machine Learning Framework

Based on the results from the previous sections and having identified the most important variables in evaluating scientific performance of the researchers, this section proposes a machine learning framework for classifying the researchers as well as predicting their performance and their deserving level of funding.

5.3.1 A Comprehensive Machine Learning Framework for Scientific Evaluation of Researchers

Funding is one of the crucial drivers of scientific activities. The increasing number of researchers and the limited financial resources has caused a tight competition among scientists to secure research funding. On the other side, it becomes even harder for funding allocation organizations to select the most proper researchers. Number of publications and citation count based indicators are the most common methods in the literature for analyzing the performance of the researchers. However, the mentioned indicators are highly correlated with the career age and reputation of the researchers since they are accumulated over time that makes it almost impossible to evaluate the performance of a researcher based on quantity and quality of his/her articles at the time of the publication. This research proposes a machine learning framework for predicting the performance of the researchers. The framework may help decision makers to better allocate the available funding to the distinguished scientists through providing fair comparative results regardless of the career age of the researchers. Our results show that the proposed framework is performing well in predicting the performance of the researchers with high accuracy as well as classifying them based on collaboration patterns, productivity, and efficiency.

5.3.1.1 Introduction

Research grants is known as one of the crucial drivers of scientific activities that can influence the size and efficiency of R&D sector and its productivity (Jacob & Lefgren, 2011). It can also affect performance of the researchers through providing them with a better access to the research resources (Lee & Bozeman, 2005). In the meantime, policies on R&D activities have evolved over the past fifty years (Elzinga & Jamison, 1995; Sanz-Menéndez & Borras, 2000). Funding agencies put a lot of efforts on selecting the best candidates for allocating grants as well as on evaluating the performance of researchers in regard to the amount of funding that they have been receiving. On the other hand, the growing number of researchers world-wide has made the competition for securing the limited financial resources even harder. For example, according to Polster (2007) the contest for receiving research funding is on the rise in Canada especially among the academic researchers mainly due to the changes in the federal funding policies, lack of university operating budgets, and

increasing research costs. The researchers' demand for funding cannot be fully satisfied by the finite financial capacity of funding agencies. However, the case could be even worse for the young researchers since the senior researchers are more known within their scientific community that might help them in getting money for research.

Peer review is the oldest measure that has been being used for evaluating researchers and their proposals. Most of the funding agencies use a committee of independent researchers to review the researchers' proposals for funding and select the most appropriate researcher(s) through a competitive process. However, the peer review process has been widely criticized in the literature due to the potential biases since accuracy of the procedure is highly dependent on the selected experts. For example, preferences of peers can affect the final decision or it can act as a gatekeeper for new research interests since peers may not come into an integrated conclusion (King, 1987). Despite the aforesaid drawbacks, the great advantage of peer review process is that the impact of the proposed research could be assessed quite easily and accurately (Allen, *et al.*, 2009). For this important reason it has still remained as one of the most popular techniques in scientific evaluation. Though, the current trend is to combine the expert review with quantitative performance indicators (Butler, 2005; Hicks, *et al.*, 2004) in order to achieve an accurate and fair evaluation since it cannot be reliable enough as a single indicator. For this purpose, citation and publication count based indicators are commonly used as the quantitative indicators of researchers' performance.

Being first introduced by Gross and Gross in 1927, citation count based indicators are commonly accepted as a proxy for the impact of a scientific publication (Gingras, 1996). In general, the mentioned metrics count the number of citations received by an article after the date that it was published and papers with higher number of citations are assumed to have higher impact. However, due to the several drawbacks of citations they are not considered by some researchers (*e.g.* Seglen, 1992) as a good measure of the quality of publications. For example, articles of famous researchers are more likely to be cited. In addition, a low quality work may receive many citations not due to its quality but because of an error in methodology or results (Okubo, 1997). However, citation counts have been widely in use as a significant index of the mean impact of a paper especially at the aggregate level (Gingras, 1996). For example, citation analyses have been used to evaluate the performance of

individual researchers (*e.g.* Garfield, 1970), quality of books (*e.g.* Nicolaisen, 2002), or performance of scientific fields and academic departments (*e.g.* Buss, 1976).

One of the reasons that scientists publish their work in the form of scientific papers is that in this way they can secure their priority in discoveries (De Bellis, 2009). According to the review of literature done by Tan (1986), in most of the cases performance evaluation of individual researchers and research departments are at least partially based on publication count measures. Due to the relatively easy access to the required data and simplicity of the calculation, publication count indicators are still widely used to analyze the productivity of the researchers or research institutes (Van Raan, 2005b). For example, publication counts have been used to a large degree for measuring the productivity of individual and departmental researchers (*e.g.* Porter & Umbach, 2001; Dundar & Lewis, 1998; Creamer, 1998, Bell & Seater, 1978). However, publication counts have also some drawbacks, *e.g.* different nature of work in various scientific disciplines (Wanner, *et al.*, 1981).

In this research we employed machine learning techniques to propose a framework for predicting the performance of the researchers as well as their deserving level of funding. The accuracy of the model was tested for a set of authors that were independent from the training data. In addition, a voting system was included in the model which weights the results of the predictors based on a randomized sample of data. This approach increases the robustness and accuracy of the model. Moreover, the accuracy of the proposed combined framework was compared with various existing data mining techniques. Hence, this paper presents an integrated highly accurate productivity prediction framework that can assist decision makers to detect the most appropriate researchers for funding allocation. The remainder of the paper proceeds as follows: Section 5.3.1.2 presents the data and methodology; Section 5.3.1.3 presents the performance evaluation results and interpretations for the proposed framework; Section 5.3.1.4 concludes; and Section 5.3.1.5 discusses the limitations and suggests directions for the future work.

5.3.1.2 Data and Methodology

5.3.1.2.1 Data

In this research we used Elsevier's Scopus to gather all the information about the NSERC funded non-student researchers. The data spans from information about the authors themselves (*e.g.* Scopus ID, their affiliation, number of publications in a given year, *etc.*) to their articles (*e.g.* year of publication, authors of the paper, keywords, *etc.*). The time interval of our research is limited to 1996 to 2010 since the data quality of Scopus was lower before 1996. The main reasons for selecting NSERC was its 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). Moreover, 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 in. SCImago was chosen for two main reasons. Firstly, it provides annual data of the journal impact factors that enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our publications database.

Moreover, we calculated several bibliometric features such as amount of funding received by an author in a given year, his/her career age, average number of co-authors, average number of publications, average number of citations, *etc.*, and stored all the calculated features in a single MySQL dataset. In addition, position of the researchers in their scientific collaboration network was evaluated by social network analysis techniques. We used Pajek software to construct the co-authorship networks of the researchers and to calculate the network structure variables at the individual level. The calculated network structure indicators were also integrated into the database. The final database contains 117,942 records. In the next section, we discuss the methodologies.

5.3.1.2.2 Methodology

We employed two types of data mining models one for classification of the researchers based on their productivity and the other one for predicting their scientific output and impact of publications as well their deserving level of funding. In this section we discuss them in detail separately.

5.3.1.2.2.1 Classification

For the classification purpose, a number of calculated bibliometric features were used as the input. They included information about quality and quantity of the publications, position of the researcher in the collaboration network, scientific discipline, and the amount of funding. The variables were calculated in a three year time window, *e.g.* for assessing the productivity of a given researcher in year 1999 his/her amount of funding was calculated from 1996 to 1998. The three-year time window for calculating the network structure variables, funding, and productivity has been already used in the literature (*e.g.* Beaudry & Allaoui, 2012). Average number of citations in a three year time window (*avgCit3*) was added to the model as a proxy for the quality of the papers. Past productivity of the researchers measured based on the average number of their papers in a three year time window was also added to the model (*noArt3*).

Three network structure indicators (*i.e.* betweenness centrality, clustering coefficient, and degree centrality) were calculated in the co-authorship network of the researchers in a three year time window. The resulted indicators were included in the model representing the impact of scientific collaboration on the productivity of the researchers. Betweenness centrality (*bc*) focuses on the role of intermediary individuals in a network and is defined for a given node *k* based on the share of times that a node *i* reaches a node *j* via the shortest path passing from the node *k* (Borgatti, 2005). Hence, researchers with high betweenness centrality in general have higher control over the researchers in the network in term of setting project priorities or knowledge diffusion. Degree centrality (*dc*) is defined based on the number of ties that a node has (*i.e.* degree of the node) in an undirected graph. Therefore, researchers with high degree centrality can be more active since they have higher number of ties to other researchers (Wasserman, 1994). In addition, in co-authorship networks degree centrality can be considered as a measure of the local team size since it is calculated based

on the number of direct connections of a researcher. Clustering Coefficient (*cc*), also called cliquishness, counts the number of triangles in the 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).

Publication and citation habits can be different in various scientific fields. For example, citing habits and the rate of citations may vary across different scientific fields in a way that in some scientific fields authors publish articles more frequently or the publications contain more references (MacRoberts & MacRoberts, 1996; Phelan, 1999). In order to stand for such variations the scientific field of the researchers was also added to the model. We performed three types of classification analysis as follows:

- To classify researchers based on their overall productivity, *i.e.* quantity and quality of the papers (Task A)
- To classify researchers according to their efficiency (Task B)
- To classify researchers based on their rate of collaboration (Task C)

The only difference in performing the above mentioned tasks was in calculating and assigning the label that is discussed in detail for each task separately. To perform Task A, a label was generated based on both quantity and quality of researchers' publications in a three year time window. For this purpose, various indicators and different weights for quantity and quality of the papers were tested. The final productivity indicator with the most robust results has three levels (*i.e.* low, normal, and high productivity) in which a relatively higher weight has been given to the quality of the papers. The same approach was taken for Tasks B and C. The efficiency of the researchers (Task B) was evaluated by calculating the cost of article indicator for each of the researchers in the database and comparing it with the average cost. The final label contains three levels representing *i.e.* low, normal, and high efficiency. For calculating researchers' collaborative behavior index (Task C), as explained earlier several combinations were tested where finally we took degree centrality and team size of the researchers in a three year time window at the individual level. This label has three levels reflecting low, normal, and high collaborative behavior of the researchers. All the labels were automatically calculated and generated by a JAVA program.

Figure 58 shows the whole process of the classification model for all the above mentioned tasks. As it can be seen, data is first preprocessed and cleaned. For this purpose, we coded several JAVA programs to check the data for redundancy, out of range values, impossible combinations, errors, and missing values and then target features were selected and data was filtered based on the records that contained all the required data. The resulted data containing all the potential features was fed into the data preparation block where at first all the features (except label) were normalized to a value between 0 and 1. This was a crucial step since the features were of different units and scales.

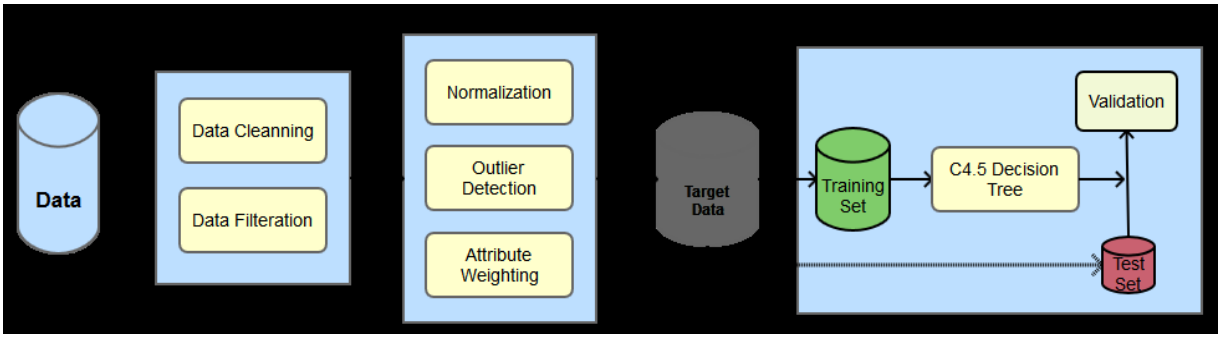


Figure 58. The classification model

Local Outlier Factor (LOF) algorithm was then used to detect the outliers. LOF that was proposed by Breunig *et al.* (2000) is based on the local density concept in which the local deviation of a given data is measured with respect to its *k* nearest neighbors. A given data is outlier if it has a substantial different density from its *k* neighbors. The final step of the data preparation step is optimizing attributes' weights. For this purpose we used an evolutionary attributes weights optimizer that employed genetic algorithm to calculate the weights of the attributes. The weighting procedure also helped us in detecting the most influential attributes. The full list of the final attributes is presented in Table 29. The resulted data was integrated into a single data repository named as the target data.

After making the data ready for the analysis, a stratified 10-fold cross validation design is used for the model validation. Cross validation is an analytics tool that is used to design and develop fine tune models. In other words, it splits data into two disjoint sets where one part is used for training and fitting a model (training set) while the other part is employed for estimating the error model (test set) (Weiss & Kulikowski, 1991). We used a nested 10-fold cross validation in which the data is split into 10 disjoint subsets in a way that the union of

the 10 folds results the original data. The method runs 10 times and in each time 1 fold is considered as the test data while the rest are regarded as the training data. For modeling the input data and performing the classification, C4.5 decision tree algorithm (Quinlan, 1993) was used where its parameters were optimized inside the validation module.

Table 29. List of attributes for the classification models

	Attribute
1	Scientific area in which the author is working
2	Total amount of funding received by each author in a 3 year time window
3	Total number of publications of each author in a 3 year time window
4	Average number of citations received by authors' articles in a 3 year time window
5	Average betweenness centrality for each author in a 3 year time window
6	Average degree centrality for each author in a 3 year time window
7	Average clustering coefficient for each author in a 3 year time window

5.3.1.2.2.2 Prediction

Figure 59 shows the general scheme of the prediction model. We used the same approach as what was already discussed in section 5.3.1.2.2.1 for the classification tasks to acquire the target data. Based on the optimized weights, we considered some extra attributes for the prediction model in comparison with the classification model. In this section, we first introduce the extra variables. In addition to the average number of citations, we used another proxy for the quality of the papers for the prediction model that is based on the impact factor of the journals in which the articles were published (*avgIf3*). Both of the mentioned measures can serve as a proxy for quality, but with a slightly different meaning. Impact factor indicates the respectability of the journal, *i.e.* the quality and the level of contribution perceived by the authors and the reviewers of the paper, whereas the citations show the impact of the article on the scientific community and on the subsequent research.

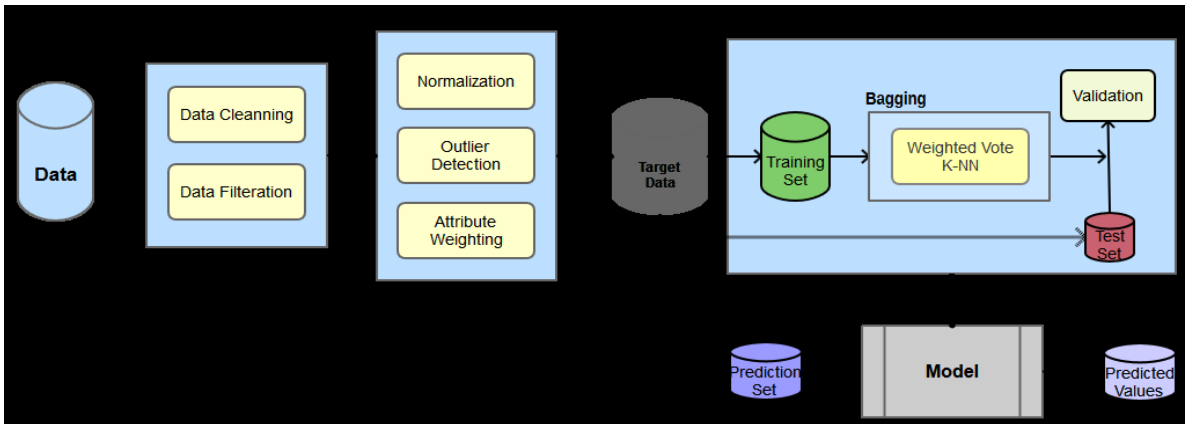


Figure 59. The prediction model

It is argued in the literature that older researchers in general can be more productive (Merton, 1973; Kyvik & Olsen, 2008) due to several reasons (*e.g.* better access to the funding and expertise sources, more established collaboration network, better access to modern equipments). Hence, the career age of the researchers (*careerAge*) was included into the model representing the time difference between the date of their first article in the database and the given year. The average number of co-authors per paper for the researchers can be counted as a measure of their scientific team size. The *teamSize* variable was also included in the prediction model. Apart from the network variables that were already discussed in section 5.3.1.2.2.1., Eigenvector Centrality (*ec*) was also added to the prediction model which is based on the idea that the importance of a node in the network also depends on the importance of his connections. Hence, an actor is more central is it is connected with other actors who are themselves central. In other words, eigenvector centrality measures how well connected an actor is in the network. Bonacich (1972) defined the eigenvector centrality of an actor based on sum of its adjacent centralities. The full list of the final attributes is presented in Table 30.

Table 30. List of attributes for the prediction models

	Attribute
1	Scientific area in which the author is working (*discip*)
2	Total amount of funding received by each author in a 3 year time window (*sumFund3*)
3	Total number of publications of each author in a 3 year time window (*noArt3*)
4	Average number of citations received by authors' articles in a 3 year time window (*avgCit3*)
5	Average impact factor of the journals in which authors' articles were published in a 3 year time window (*avgIf3*)
6	Average betweenness centrality for each author in a 3 year time window (*btwn3*)
7	Average degree centrality for each author in a 3 year time window (*deg3*)
8	Average clustering coefficient for each author in a 3 year time window (*clust3*)
9	Average eigenvector centrality for each author in a 3 year time window (*eigen3*)
10	Average scientific team size of the researcher (*teamSize*)
11	Career age of the researcher (*careerAge*)

We predict two target variables as follows:

- The number of publications of a given researcher (Task 1)
- The amount of funding that a given researcher deserves (Task 2)

To perform Task 1, we considered the number of publications of the researchers as the label while the label for Task 2 was the amount of funding. As mentioned earlier, the procedure for preparing the target data is similar to the classification model. The difference

is in the algorithm where in the prediction model we used ensemble meta-algorithm to improve the accuracy of the prediction. For this purpose, bootstrap aggregating (bagging) approach was employed. Bagging is an ensemble method that makes random subsets of the data and trains them separately where the final result is obtained by averaging over the results of the separated models (Breiman, 1996). Bagging is a nested module in which we used weighted vote 10-Nearest Neighbor (10-NN) algorithm to train the data and create the model. In weighted vote 10-NN the distance of the neighbors to the given data is considered as a weight in the prediction in a way that neighbors that are closer to the given data get higher weights. To train and build the model, data in the range of 1996 to 2009 was used. A separate disjoint data for 2010 (prediction set) was used for testing the accuracy of the prediction model. The final result of the prediction model for Task 1 is the predicted number of publications for the researchers in the prediction set while for Task 2 the output of the model is the amount of funding that researchers deserve to receive in the given year. In the next section, the results of the discussed models are presented.

5.3.1.3 Results

5.3.1.3.1 Classification

The proposed framework was fed with the data that was already explained in section 5.3.1.2.1 to evaluate its accuracy for the three defined classification tasks (Task A, B, and C). Moreover, we separately tested several machine learning algorithms to be able to compare the accuracy of the proposed classification framework (*PCF*) with some well-known classifiers. For this purpose, we listed the results for the top three most accurate algorithms for each task along with the one for the framework. Models were trained on the data from 1996 to 2010. Figure 60 shows the results for Task A. As it can be seen the accuracy of PCF in Task A is reasonably higher than the other algorithms. To compare the accuracy more accurately we evaluated the confusion matrices of PCF and 10-NN which has the nearest accuracy to PCF.

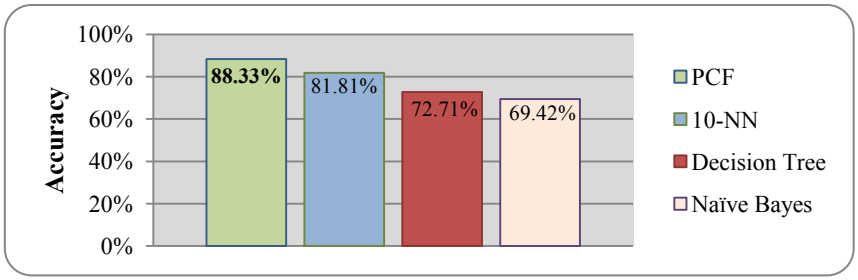


Figure 60. Accuracy of PCF vs. selected algorithms, Task A

Confusion matrix was introduced by Kohavi and Provost (1998) and shows the actual and predicted classifications done by a classifier and is used to evaluate the performance of the classification system. *Precision* and *recall* are two of the measures that are used in the confusion matrix. According to the definition, precision is the proportion of the total number of correct predictions. Recall of a label in a multi-class problem is defined as the ratio of the correctly predicted cases for that class over the total number of predictions. As it can be seen in Table 31, Although PCF and 10-NN precision and recall is almost comparable for the predicted high and true low cases, PCF has higher rates of precision and recall in all the sub-classes. Hence, PCF is a more accurate classifier for the subject problem since it provides higher precision and recall rates. Lift chart[29] of PCF for Task A is presented in Appendix C.

Table 31. Confusion matrix of PCF vs. 10-NN, Task A

		PCF	10-NN
Precision	Predicted Low	94.28%	87.74%
	Predicted Normal	78.51%	67.61%
	Predicted High	84.59%	83.79%
Recall	True Low	94.36%	92.88%
	True Normal	79.69%	67.78%
	True High	82.53%	68.53%

The same analysis was done for evaluating the performance of PCF in classifying the data for Task B. As it can be seen in Figure 61, accuracy of PCF is higher than the other top three most accurate algorithms. Interestingly, apart from 10-NN other classifiers (Naïve Bayes and Decision Tree) have considerably lower accuracy than PCF. Although Naïve

[29] Lift chart is another tool to see the performance and the predictive power of a model.

Bayes algorithm is simple and computationally efficient, it is based on strong attribute independence assumptions which might be one of the reasons that this algorithm is not working well for Task B classification. Decision trees are also simple and very easy to understand. However, apart from the cost of operation and its complexity there are some concepts that decision trees cannot learn them. Moreover, since our problem is a multi-label classification, the information gain in decision tree can be biased in favor of attributes with more number of observations (Deng, *et al.*, 2011) hence the algorithm might not be able to model the data accurately.

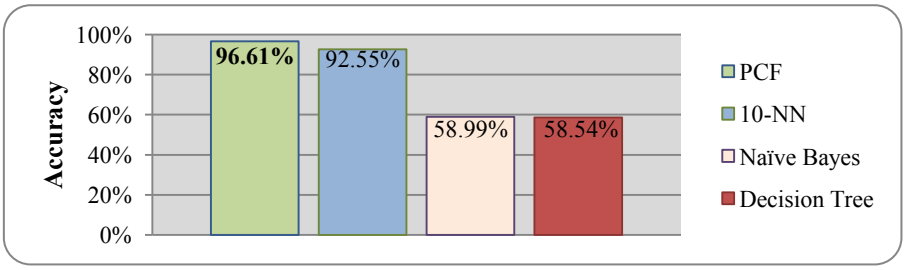


Figure 61. Accuracy of PCF vs. selected well-known algorithms, Task B

We took 10-NN (as it had the closest accuracy to PCF) and compared its confusion matrix with PCF. As it can be seen in Table 32, precision and recall rates for PCF is higher than the ones for 10-NN except for the precision of the predicted high category for which 10-NN is slightly higher. The high accuracy of 10-NN is not very surprising since these classifiers work well when the size of the training data is large. In addition, in our case we have several features which 10-NN can benefit from to characterize each label based on multiple combinations of the attributes and increase the accuracy. The lift chart for Task B is presented in Appendix C.

Table 32. Confusion matrix of PCF vs. 10-NN, Task B

		PCF	10-NN
Precision	Predicted Low	98.03%	92.99%
	Predicted Normal	94.78%	90.11%
	Predicted High	96.58%	97.14%
Recall	True Low	97.75%	96.53%
	True Normal	95.89%	89.72%
	True High	94.85%	87.06%

In the last part of this section we assess the performance of PCF in classifying researchers based on their collaboration patterns (Task C). According to Figure 62, the proposed framework performs better than other available algorithms in performing Task C with 98.90% of accuracy. Decision Tree is next in terms of accuracy while 10-NN and Naïve Bayes are coming after respectively. Analysis of the confusion matrix (Table 33) reveals that PCF has higher rate of precision than Decision Tree except for the predicted low category where the difference is almost negligible (99.55% vs. 100%). For the recall rates PCF also performs better except for the true high category where the difference is small (96.90% vs. 98.70%). The lift chart for Task C is presented in Appendix C.

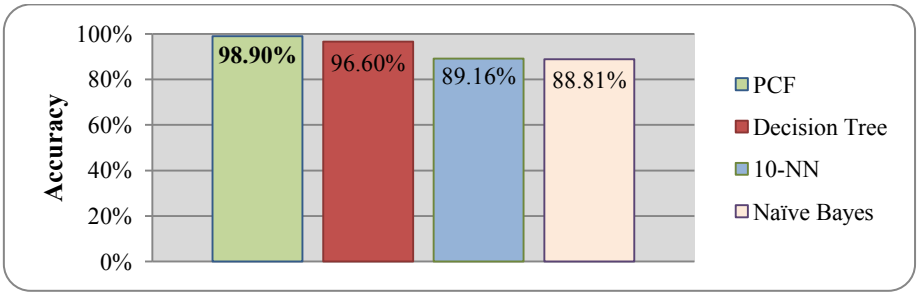


Figure 62. Accuracy of PCF vs. selected well-known algorithms, Task C

As it was observed in this section, PCF performs reasonably well in classifying the researchers based on various measures such as collaboration patterns, productivity, and efficiency. In the next section, we check the performance of the proposed prediction framework (PPF) in predicting the number of publications (Task 1) and amount of funding (Task 2) that researchers deserve to receive.

Table 33. Confusion matrix of PCF vs. Decision Tree, Task B

		PCF	Decision Tree
Precision	Predicted Low	99.55%	100%
	Predicted Normal	98.32%	92.77%
	Predicted High	97.17%	91.48%
Recall	True Low	99.70%	96.15%
	True Normal	98.16%	96.66%
	True High	96.90%	98.70%

5.3.1.3.2 Prediction

In this section we present the results of the performance evaluation of the proposed prediction framework (PPF) in predicting productivity of the researchers (Task 1) as well as the amount of funding that they deserve in a given year (Task 2). For this purpose we trained the model with the data from 1996 and 2009. Disjoint 2010 data was fed into the learned model to predict the target variables. We also tested the accuracy of the model with several well-known machine learning algorithms. We list the accuracy of PPF along with two other algorithms that showed the highest accuracy in predicting the target variable in each of the defined tasks.

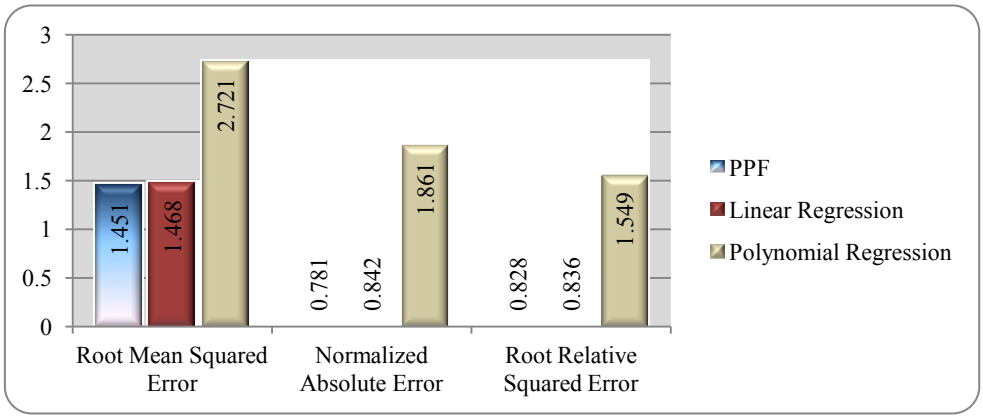


Figure 63. Accuracy of PPF vs. selected well-known algorithms, Task 1

Figure 63 shows Task 1 prediction errors for PPF, linear regression, and polynomial regression of degree 3. We considered three error measures for comparing the performance of the mentioned algorithms. Root mean squared error is one of the main measures for comparing the accuracy of the prediction models and is defined as the square root of the average of the squares of errors. According to Figure 63, PPF is predicating the number of publications of the researchers with 1.451 average deviation between the predicted value and the real number of publications. Normalized absolute error is the absolute error (difference between the predicted value and the real value) divided by the error made if the average would have been predicted. The root relative squared error takes the average of the actual values as a simple predictor to calculate the total squared error. The result is then normalized by dividing it by the total squared error of the simple predictor and square root is taken to transform it to the same dimension as the predicted value. As it can be seen PPF is

performing better in all the three measures where the degree 3 polynomial fit is the worst. A sample of the prediction results is presented in Appendix C.

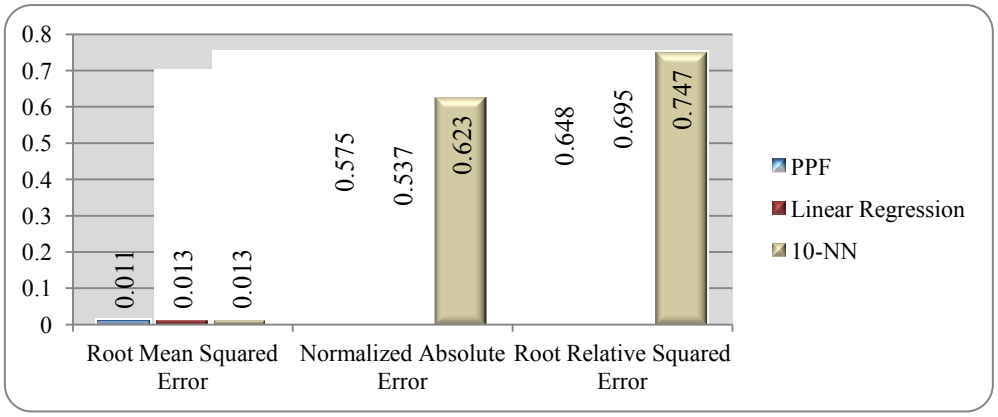


Figure 64. Accuracy of PPF vs. selected well-known algorithms, Task 2

We did the same analysis for comparing the accuracy of PPF with other selected highly-accurate algorithms in predicting the amount of funding that a researcher deserves to receive (Task 2). In performing Task 2, linear regression and 10-NN algorithms were the two closest algorithms to PPF in terms of the prediction errors. According to Figure 64, root mean squared error of PPF is the lowest were the other two algorithms are doing the same with a slightly higher error than PPF. Although linear regression normalized absolute error is a bit lower than PPF, its root relative squared error surpasses PPF. Hence, according to the results, it can be seen that the overall performance of PPF is slightly better than the other available algorithms.

5.3.1.4 Conclusion

In this paper we used bibliometric indicators as well as social network analysis features to classify researchers based on their collaboration patterns, productivity, and efficiency. We also proposed a model to predict the number of publications of researchers along with their competence level for receiving grants. According to our results it is feasible to employ machine learning algorithms for classification of the researchers based on various criteria. Moreover, it was shown that the proposed framework can predict the productivity and deserving level of funding of the researchers with relatively high accuracy. As it was shown, even though in some minor cases the other algorithms perform slightly better than the proposed framework in a single task, when we consider all the defined tasks the performance

of the proposed framework is much higher than the other algorithms. In addition, the unique procedure that was presented in this research highlighted the most important features in classifying researchers and predicting their performance.

Although some researchers recently worked on citation prediction using machine learning algorithms (*e.g.* Fu & Aliferis, 2010; Lokker, *et al.*, 2008) to our knowledge this is the first study that focused on productivity and competence level of funding prediction as well as classifying researchers using bibliometric and social network analysis indicators. In addition, we used attribute weighting to rank the features based on their importance and employed outlier detection to filter the data. Hence, the intensive preprocessing stage along with attribute selection procedure helped the model to achieve high predictive power and accuracy rate. The result of attribute weighting module also shed light on the influential attributes in predicting or categorizing the target researchers. Moreover, several features of similar nature were employed in the model to reinforce its accuracy. For example, we used average number of citations and average impact factor of the journals to represent the quality of the publications. Another example is the degree centrality and scientific team size. These attributes of similar nature surely empowered the accuracy of the model by providing it with more dimension.

To conclude, our results show that it is feasible to design and use classification and prediction tools to evaluate different aspects of scientific activities of researchers. It is obvious that peer reviewing cannot be completely replaced by such tools. The proposed frameworks in this research can help decision makers in setting both long-run and short-term strategies in regard to the funding allocation and/or analyzing researchers' productivity and scientific collaboration patterns among the researchers. In addition, since our framework uses high dimensional data and a large dataset spanning from 1996 to 2010 to learn the model the result is not created based on limited criteria or data. Therefore, it can also help decision makers to establish a fairer funding allocation or scientific evaluation system.

5.3.1.5 Limitations and Future Work

The first limitation was in regard to the source of data for which Scopus was selected. Since Scopus and other similar databases are English biased, hence, non-English articles are

underrepresented (Okubo, 1997). 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.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among researchers as we were unable to capture other links that might exist among researchers like informal relationships. These types of connections are never recorded and thus cannot be quantified, but there are certainly some knowledge exchanges occurring during such associations that could affect the network performance. In addition, there are also some drawbacks in using co-authorship as an indicator of scientific collaboration since collaboration does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997). Hence, future work can address this issue by taking other types of collaboration networks into the consideration.

For assessing the quality of the papers based on citation count we did not account for self citations, negative citations, or special inter-citation patterns among a number of researchers. Although we also used another proxy (average impact factor of journals) to overcome this limitation, it can be addressed in the future works as well.

5.4 Survey Data Analysis

We made several assumptions in this thesis. For validating the assumptions we held 30 interviews with the selected researchers from the database. The researchers were selected using stratified sampling method. In addition, a questionnaire was designed and sent to 8,000 researchers. In this section the results of the survey data analysis is presented.

5.4.1 Funding, Collaboration, and Scientific Performance: A Survey Analysis

Research is highly dependent on funding. Growing number of researchers and the limited funding resources has caused a severe competition among researchers to secure their required resources. On the other hand, the complex and interdisciplinary nature of the modern science has encouraged researchers to collaborate more. Hence, apart from financial resources finding right partners to collaborate has become also important. Based on the data from a questionnaire study we found that researchers consider different criteria in selecting

their collaboration partners based on their available level of funding and the nature of collaboration. In addition, the results suggest a positive relation between funding and number of publications.

5.4.1.1 Introduction

According to the *sacred spark* hypothesis the differences in researchers' productivity can be mainly due to the predefined differences in the characteristics of the researchers that cause diverse range of personal capabilities and level of motivation to solve a research problem (Cole & Cole, 1973). However, several studies criticized the mentioned hypothesis as it does not provide a concrete and comprehensive explanation for the diversities in the performance of the researchers (Allison & Stewart, 1974; Fox, 1983; Stephan, 1996). In addition, there is no evidence that the differences in the rate of publications among researchers exactly comply with their capabilities. Moreover, even if we suppose that the characteristic varieties among scientists can partially explain their performance, it is hard to justify different performance of a same researcher during various stages of his/her career (Stephan, 1996). Hence, other external factors influence scientific activities and performance of researchers.

Although governments in many western countries invest on research and development activities to secure their world-wide competitive position, they have been always looking for ways to fulfill society's requirement with lower money. One reason could be the limitedness of the financial resources and the fact that the number of applicants is also growing. In addition, governments are under public pressure to cut the share of taxpayer generated money (Liefner, 2003). On the other hand, research grant is known as one of the crucial drivers of scientific activities (Jacob & Lefgren, 2011) since it can affect the performance of the researchers through providing them with a better access to the research resources (Lee & Bozeman, 2005). Moreover, funding can influence the scientific collaboration patterns among researchers that might result in higher productivity. Hence, due to the mentioned factors governments and funding agencies not only aim for selecting the most potential and suited candidates for funding allocation they also employ performance assessment methods to evaluate the outcome of the funded research.

Nature of the modern science has become more interdisciplinary, complex, and costly than before that force researchers to collaborate more (Lee & Bozeman, 2005). In addition, the limited research resources may encourage researchers to get involve more in collaborative research. Hence, it is normal that researchers tend to collaborate as part of their scientific activities (Beaver & Rosen, 1979) due to several good reasons such as more efficient use of resources or getting access to expensive equipments (Thorsteinsdóttir, 2000). However, finding right partners to collaborate and coordination costs of working with others can act as barriers in scientific collaboration (Landry & Amara, 1998). Bozeman and Corley (2004) found that especially senior researchers do not highly tend in collaborative activities where they prefer more to mentor research students. On the other hand, senior researchers on average benefit from more established professional networks that might help them to increase their productivity. Therefore, despite the implicit assumption of the positive relation between collaboration and productivity in various studies (*e.g.* Lotka, 1926; Zuckerman, 1967; Godin & Gingras, 2000), the relation between them is not clear (Lee & Bozeman, 2005).

Universities, colleges, and research institutes are considered as the main players in the process of knowledge production and diffusion (Gulbrandsen & Smeby, 2005). Hence, analyzing the interrelations among funding, scientific collaboration, and scientific performance within the academia environment can be informative. Although several empirical studies showed that funding has a positive impact on the productivity of the researchers at the individual level (Lee & Bozeman, 2005; Stephan, 1996), the intensity of the effect varies. According to the literature several factors can influence the intensity of the impact of funding on scientific productivity, *e.g.* the career age of the researchers (Arora & Gambardella, 1998) or the amount of funding (Godin, 2003). However, there are also some studies that found no significant relation between funding and number of publications (*e.g.* Gaughan & Bozeman, 2002; Huffman & Evanson, 2005). Although it is stated that funding can positively influence scientific collaboration (Adams, *et al.*, 2005; Arora & Gambardella, 1998; Katz & Martin, 1997), the impact of funded collaboration on scientific output is less obvious (Defazio, *et al.*, 2009). However, according to Arora and Gambardella (1998) scientific collaboration highly depends on the existence of funding in a way that securing new financial resources can encourage researchers to collaborate.

Policy makers often emphasize on the direct interaction between universities and industry as the main driver of technology based economic development of the country. The nature of such collaboration is indirect since universities train skillful graduates to be involved in the industry. However, increased direct collaboration between the academic environment and industry can enhance knowledge production and diffusion (Gibbons *et al.*, 1994; Martin, 2003; Martin & Etzkowitz, 2000). Direct commercialization of the research output of academic researchers is one of the examples of the policies that the decision makers set for fostering the mentioned collaboration (Godin & Gingras, 2000; Van Looy, *et al.*, 2004). There exists opposing and supportive studies in the literature for the collaboration between academia and industry. Increased pressure on the academic researchers, conflicts between the open science nature of academic environment and competitive commercialized nature of industry, decline in the teaching task of the professors are examples of the negative impact of the increased collaboration between industry and universities (Geuna, 2001; Geuna & Nesta, 2003; Vavakova, 1998). In addition, there exist barriers (*e.g.* intellectual property) on partnering universities with industry (Hall, *et al.*, 2001) that make the collaboration even harder. On the other hand, it has been argued that having closer relations with industry can strengthen universities through bringing more autonomy and flexibility to the academic researchers (Kleinman & Vallas, 2001) while leading them to becoming *entrepreneurial institutions* (Clark & Clark, 1998; Etzkowitz, 2003). The remainder of the paper proceeds as follows: Section 5.4.1.2 presents the main propositions of this research; Section 5.4.1.3 represents the data and methodology; Section 5.4.1.4 presents the empirical results and interpretations; Section 5.4.1.5 concludes; and Section 5.4.1.6 discusses the limitations of the research and suggests directions for the future work.

5.4.1.2 Conceptual Framework

With regards to the general increasing emphasis on benefits of the funded research for the society, this paper aims to analyze the interrelation among funding, collaboration, and scientific profiles of the researchers. The fundamental hypothesis of this article is that following different collaboration patterns by researchers (*e.g.* in various scientific disciplines) along with the fact that they have different scientific profiles significantly affect their research budget and productivity. Securing funding is a crucial factor for scientific

activities as lack of money can limit the research opportunities (Gulbrandsen & Smeby, 2005) and decrease the productivity. In this section, we present the theoretical scope of the research and discuss the hypotheses.

5.4.1.2.1 Research Funding

The importance of securing research funding, as one of the main determinants of scientific activities, is growing within the scientific communities. Limited sources of funding and the increasing number of applicants are two of the main reasons that have made the competition for getting research money tighter than ever. Apart from the important role of funding in stimulating scientific activities, they play a significant role in researchers' academic career which is partly due to the fact that some universities expect their faculty members to financially contribute to the university (Polster, 2007).

Several factors can influence the amount of funding that a researcher receives. It is argued in the literature that the level of funding is related to the academic excellence and reputation of the researcher (Polster, 2007). According to Arora and Gambardella (1998) past productivity of the researchers positively affects their future amount of funding. The amount of research funding in the past can also influence the level of funding in the subsequent years (Beaudry & Allaoui, 2012). Moreover, role of the researcher in the collaboration network can also affect their productivity (Beaudry & Allaoui, 2012). On the other hand, higher amount of funding might have a positive impact on scientific productivity (Jacob & Lefgren, 2011; Payne & Siow, 2003; Zucker, *et al.*, 2007), or in some cases it may affect the scientific output negatively (*e.g.* Huffman & Evenson, 2005).

Hypothesis 1 becomes: *Past productivity and collaboration patterns of the researchers positively affects their level of funding*.

5.4.1.2.2 Scientific Output

Number of publications is widely used in the literature as a proxy of scientific productivity. It is easy to assess the research impact of highly productive and Nobel laureates (Hirsch, 2005). However, prolific researchers may benefit from the extra amount of funding available to them to cover the research expenses (*e.g.* materials, equipment) and increase their productivity. This is termed as the *credibility cycle* where eminent researchers may

acquire disproportionate amount of credit and resources (Latour & Woolgar, 1986). Apart from funding other factors (*e.g.* demographics) can also influence the productivity of researchers. Hence we suggest that *profile of the researchers and amount of funding can influence their scientific productivity* (**Hypothesis 2**).

5.4.1.2.3 Scientific Collaboration Patterns

The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006). The complex nature of modern science encourages researchers to be more collaborative (Lee & Bozeman, 2005). Being involved in larger scientific teams enables researchers to benefit from various expertises in order to increase the quality of their work. Moreover, they can increase the team efficiency though economies of scale. In addition, tasks can be divided among the team members where by means of a good coordination the overall productivity can be also increased. Hence, collaboration patterns among scientists play an important role in scientific activities. Godin (1998) focused on Canadian academics and found that researchers who are collaborating with industry produce on average more articles than their counterparts without such collaboration.

In general, it can be assumed that researchers who have more funding form larger scientific teams and collaborate more with the other scientists since more money might enable them to overcome the collaboration obstacles (*e.g.* coordination costs among the team members) easier. In other words, financial investment can change the structure of research groups and affect the collaboration among the scientists. Hence, *collaboration motives vary among the researchers based on their level of funding available* (**Hypothesis 3**). Moreover, researchers in different scientific fields do not necessarily follow the same approach in collaborating with their colleagues. For example, it is argued in the literature that researchers who are involved in more applied research tend to collaborate more with other scientists who are inside or even outside of their scientific community (Ernø-Kjølhede, *et al.*, 2001; Katz & Allen, 1982).

Establishing an effective link between university as one of the drivers of high-technology-based scientific development and industry is essential for a successful knowledge diffusion and innovation system. The most commonly stated advantages of the

collaboration between academia and industry for a firm include having access to knowledgeable researchers and well-educated graduates, university facilities, new scientific knowledge, state-of-the-art information, and obtaining cost-effective solutions to technical and R&D problems (Wang & Shapira, 2012). In fact, it is argued that in the absence of the academic research there would be substantial delays and much higher costs, which would often make the new product development economically undesirable (Mansfield, 1995). We suggest that, *collaboration motives vary for finding academic and industrial collaborators* (**Hypothesis 4**).

5.4.1.3 Data and Methodology

We focused on the researchers who received grants from NSERC during the period of 1996 to 2010. NSERC was selected since it is the main federal funding organization in Canada, and almost all the Canadian researchers in natural sciences and engineering receive a research grant from NSERC annually (Godin, 2003). The data of the NSERC funded researchers was extracted for the mentioned time interval that resulted in 75,967 records of distinct researchers. As the next step, we gathered all the articles that were published by our target funded researchers for the period of 1996 to 2010. The articles information were collected form Elsevier's Scopus. We decided to focus on the period of 1996 to 2010 since the data quality of Scopus was lower before 1996.

Using bibliometric indicators, statistical analyses and social network analysis techniques, we analyzed the collected data to distinguish nine groups of researchers, *i.e.* high funding, low funding, average funding, most productive, least productive, normal productivity, most collaborative, least collaborative, and average collaborative researchers. We selected these nine groups of respondents in order not to be biased in direction of any specific types of the researchers (*e.g.* elite academic researchers). A questionnaire study was designed by a team of statistical analysis experts and reviewed by a number of selected peers in different scientific fields. The questionnaire contains four parts addressing research, collaboration, and funding profiles of the researchers, and specifically analyzing the barriers and motives for scientific collaboration. Hence, three central background variables are addressed in this paper *i.e.* funding, scientific output, and collaboration.

We distinguished researchers based on several items *i.e.* gender, age range, province, language, and scientific position. The prepared questionnaire was sent to 4,000 of the target funded researchers where the response rate was 4.9% resulting in total 196 responses. The respondents were selected by stratified sampling method. Interestingly, researchers who were more productive and more collaborative responded more to the questionnaire. This is in line with findings of Kyvik (1991). However, we continued collecting responses to have enough responses from all the nine predefined groups of researchers. To compare the scientific performance and collaboration of the NSERC funded researchers with the highest scientific standards, we also sent a similar questionnaire to the researchers affiliated with the top 10 world high ranking universities in 2013. Using stratified sampling method, the questionnaire was forwarded to 4,000 of the target researchers from which we received 205 valid responses.

After collecting the survey data, we used survey data analysis technique to assess the impact of the influencing factors on the target variables. For this purpose, several bi-variate relationships between independent and dependent variables were tested and proportions of the variables were analyzed. As the final stage, we used regression analysis to evaluate the effect of the defined factors on target variables. All the proportion analyses results that are presented in this paper are significant at the level of at least 95%. In the next section, we will first present descriptive analysis of some of the important indicators. The section will continue by reporting the results of the statistical analyses.

5.4.1.4 Results

5.4.1.4.1 Descriptive Statistics, Funding vs. Scientific Output

Before turning to the statistical analysis we first briefly describe some of the explanatory variables. All the results are significant at the level of 95%. Two major categories of indicators are used for evaluating the quality of the papers, one is based on the number of citations received by an article and the other one is based on the impact factor of the journal in which the article was published. It is argued in the literature that journal impact factor cannot be considered as a good paper quality measure since it is highly discipline dependent and editorial policies can also affect the impact factor (Moed & van Leeuwen, 1996; Seglen, 1997). Advent of the digital age in 90s facilitated access to the publications, hence

weakening the traditional bound of the papers to their journals since papers can be read and cited according to their own quality and worthiness. Lozano *et al.* (2012) evaluated the relation between actual citation counts and journal impact factors during the period of 1902 to 2009. Interestingly they found that after 1990 the relation between number of citations and journal impact factor has been weakening. Although number of citations has also some drawbacks (*e.g.* negative citations and self-citation), citation based indicators are considered as the common practice in measuring the overall impact of an article (Seglen, 1992).

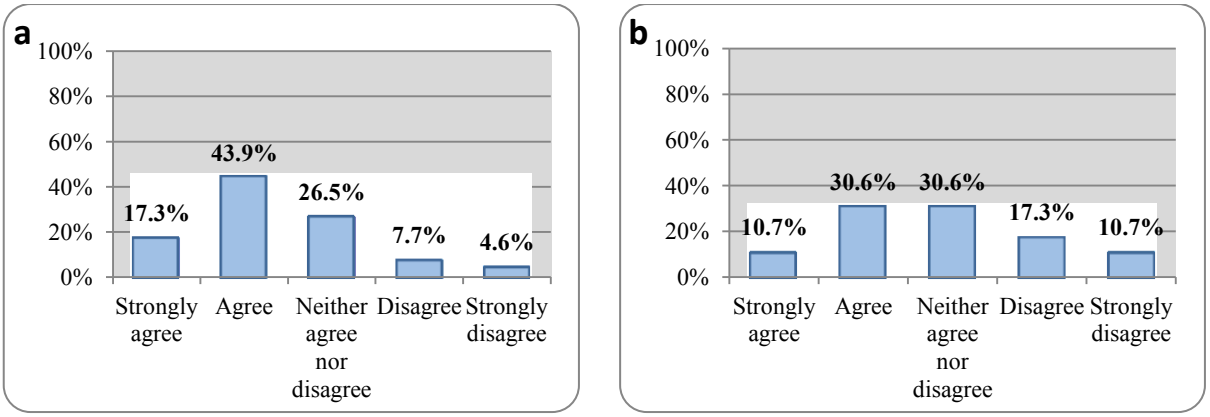


Figure 65. a) Citation count as a measure of paper quality, NSERC funded researchers, b) Journal impact factor as a measure of paper quality, NSERC funded researchers

Our results also confirm the higher credibility of the citation based indicators in comparison with the journal impact factor. As it can be seen in Figure 65, 61.2% of our respondents voted to the validity of the citation count as a proxy of paper quality while just 41.3% of them agreed that journal impact factor is a good representative of the publications' quality. According to Figure 66-a, the majority of the top 10 universities' researchers also agreed that the citation counts can be a good proxy for the quality of the papers, although the percentage is a bit lower than the ones for the NSERC researchers. Figure 66-b also confirms this finding since more than 40% of the respondents affiliated with the top 10 universities did not believe in the journal impact factor as a good measure of publications' quality.

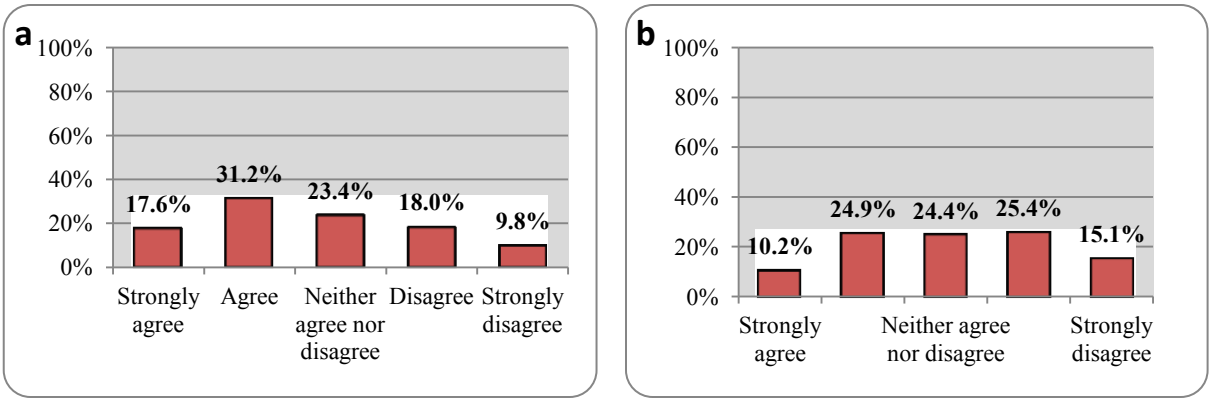


Figure 66. a) Citation count as a measure of paper quality, top 10 universities' researchers, b) Journal impact factor as a measure of paper quality, top 10 universities' researchers

Funding agencies and organizations measure the performance of the grantees in regard to the amount of funding that they have been receiving. However, evaluating the relation between the output of researchers in terms of quantity and quality of publications with their level of funding has been a challenging issue for the policy makers. Although in most of the cases a positive relation has been observed between funding and productivity (*e.g.* Payne & Siow, 2003; Jacob & Lefgren, 2007), there also exist some studies that found no relation (*e.g.* Carayol & Matt, 2006) or negative relation (*e.g.* Huffman & Evenson, 2005). One reason for the inconsistent results could be different scope, area, and datasets of the studies.

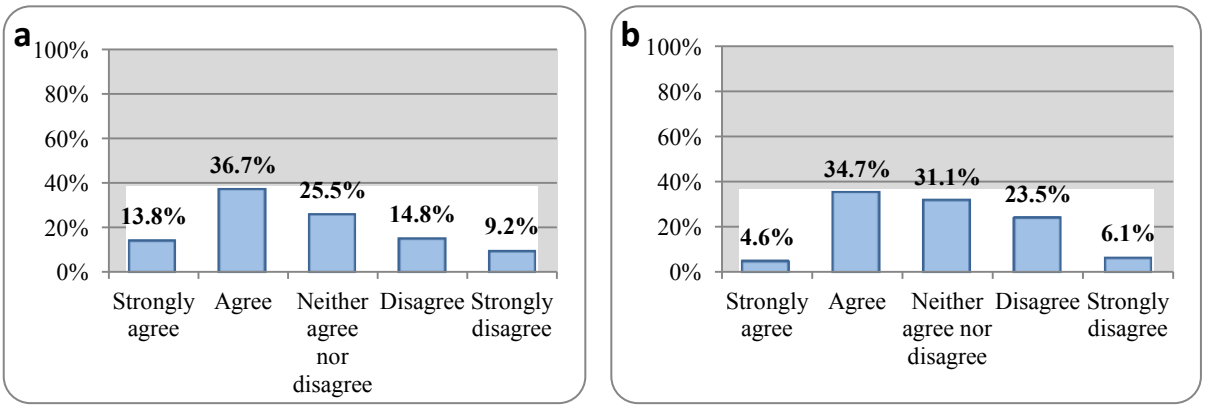


Figure 67. a) Higher funding result in higher number of publications, NSERC funded researchers, b) Higher funding result in higher quality papers, NSERC funded researchers

The other interesting point is that even in the cases with the positive relation, the direction and intensity of the relation is time dependent. In other words, as the time passes the average productivity of the researchers may not exactly follow their average level of funding. Examples are Zuckerman (1996) who observed that the productivity of the Nobel

laureates decreases after winning the Nobel Prize, or Lee and Bozeman (2005) who found that after a certain age the productivity of researchers declines.

As it can be seen in Figure 67, majority of the respondents believe that higher amount of funding enables them to not only increase the rate of publications but also to improve the quality of their work. However, researchers from the top 10 high ranking universities world-wide did not respond exactly the same as the NSERC funded researchers. According to the top 10 universities' researchers (Figure 68), there might exists a positive relation between funding and the number of publications but interestingly they believe that higher level of funding does not necessarily result in higher quality of work. This can be due to the reason that they tend to produce high quality works by default in order to secure or improve their academic profile and position and their papers would be on average of higher quality in comparison with the other researchers. Hence, no matter of the level of funding top universities' researchers may maintain the quality level of their publications. In the next section, we highlight the most important motives and barriers in scientific collaboration.

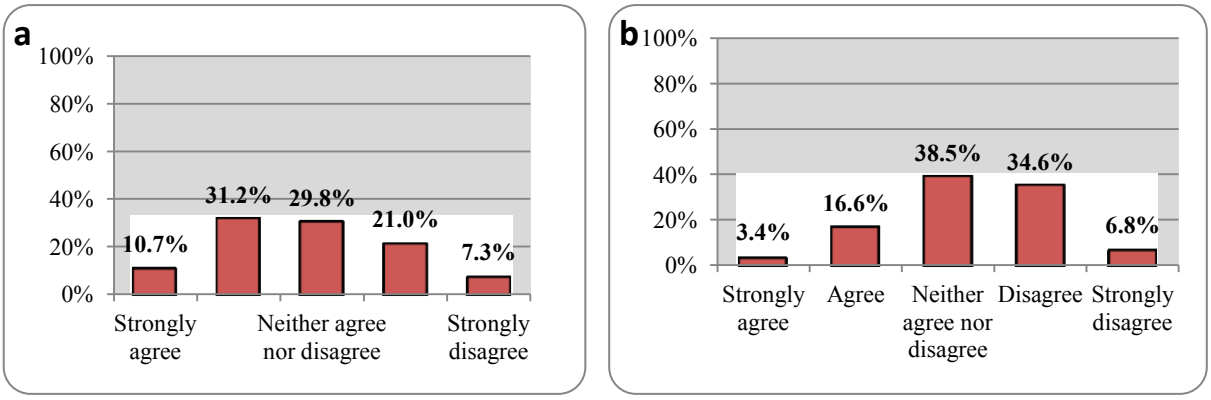


Figure 68. a) Higher funding result in higher number of publications, top 10 universities' researchers, b) Higher funding result in higher quality papers, top 10 universities' researchers

5.4.1.4.2 Descriptive Statistics, Scientific Collaboration

Scientific collaboration is defined as the process through which researchers with a common goal work together to produce new scientific knowledge (Katz & Martin, 1997). The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006). As the nature of the modern science is more costly, complex, and interdisciplinary researchers tend more to get involved in collaborative research (Lee & Bozeman, 2005). In an early study, Beaver and Rosen (1978) listed several important

motives for collaboration, *e.g.* better access to expertise, skills, equipments, materials, increasing productivity. Beaver (2001) added a new important personal reason for collaboration, *i.e.* fun, amusement, and pleasure. Hence, apart from the professional factors that matter in forming the scientific collaboration researchers also consider personal relations and feelings in collaborating with other scientists.

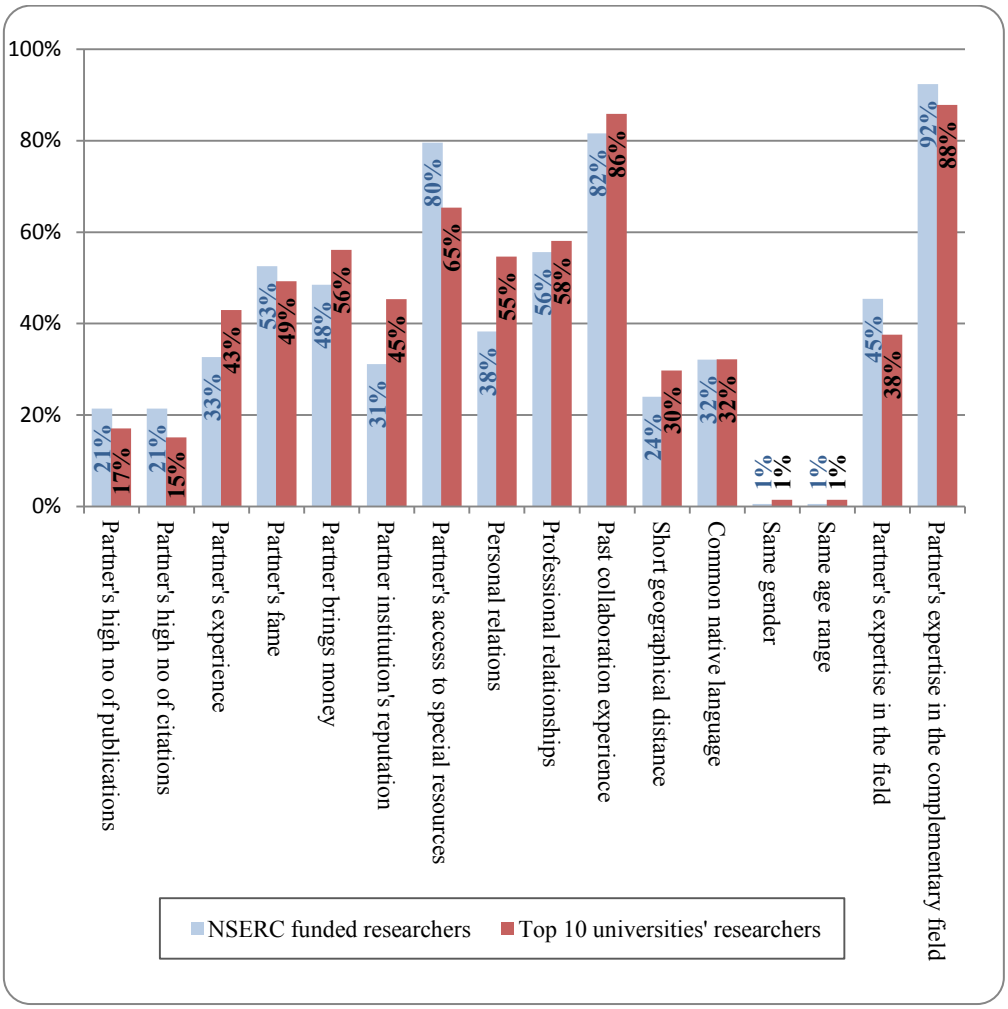


Figure 69. Motives for scientific collaboration

We checked for 16 different motives for collaboration in our survey. As it can be seen in Figure 69, there is not much difference in the opinions of the NSERC researchers and the top 10 universities' researchers. Both groups mentioned the expertise of the potential partner in the complementary field of research and the past collaboration experience as the most important reasons in collaborating with other scientists while indicating gender and age range of the partner as the least important factors. The highest difference in the results from

the two groups of the researchers is for the access to the special resources where the top 10 universities' researchers voted less (65%) in comparison with the other group (80%). This is quite reasonable since researchers of the high ranking universities may have better access to special equipments and resources in comparison with the other researchers. Interestingly, partner's productivity in terms of quantity and quality of publications just attracted around 20% of the votes. Hence the data from the survey shows that professional and personal relations along with partner's expertise play more important role in collaborating with other researchers.

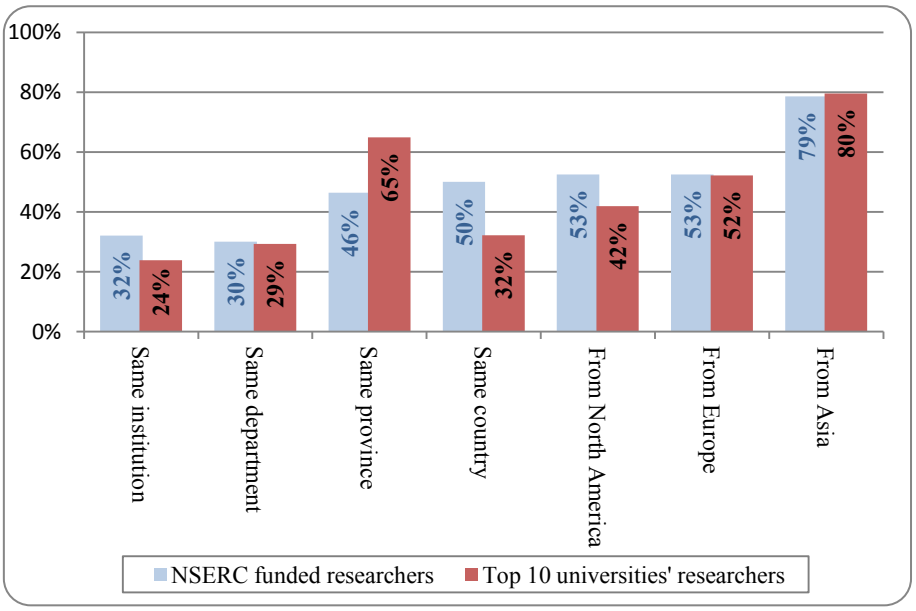


Figure 70. Collaboration disfavor, geographical distance and location

We further investigated the impact of distance and geographical location in selecting collaboration partners. As it can be seen in Figure 70, both groups of our respondents tend less to collaborate with Asian and European researchers but interestingly prefer to collaborate with researchers who are located in the same country as they are. Hence, it seems that although the invention of the digital age has lowered the distances among researchers and provided them with a better access to the facilities and resources, the physical distance still plays an important role among researchers for selecting their partners. Apart from the physical distance, cultural issues can also influence their decision since various cultures might increase the coordination costs that will result in a less effective collaboration. Moreover, as it can be observed in the figure they prefer to collaborate more to their colleagues who work in the same institution and/or department. In the next section, we will

statistically analyze the impact of the influencing factors on researchers' productivity, collaboration, and funding.

5.4.1.4.3 Statistical Analysis

In this section, we first statistically analyze the impact of productivity on the amount of the research budget. The section will continue with evaluating the impact of collaboration motives on the size of the academic and industrial collaboration networks of the researchers. Table 34 shows the results of the two-way analysis between the research budget and the number of publications for the NSERC funded researchers. As it can be seen the majority proportion of the publications belong to the high funding group of researchers (>$300K). In addition, within the high funding group most of the researchers have produced more than 100 articles where it is not the case for the other categories of the funding level. Interestingly the lowest funding group (<$40K) is not producing the least number of publications. One reason could be the involvement of young researchers and professors in this group that are eager to improve their scientific and academic position both in their institute and within the scientific community.

Table 34. Research budget vs. number of publications, NSERC funded researchers

Research budget	**Number of publications**								
	<10	10-19	20-29	30-39	40-49	50-69	70-100	>100	Total
< $40K	.0408	0	.0255	.0204	0	.0306	.0102	.0714	.199
$40K-$80K	.0102	0	.0102	.0102	.0153	.0204	.0306	.0561	.1531
$80-$150K	.0153	.0102	0	.0153	.0153	.0153	.0612	.0816	.2143
$150K-$300K	.0102	.0051	.0102	.0051	0	.0153	.0051	.0816	.1327
>$300K	.0153	.0102	0	.0102	0	.0255	.0102	.2092	.2806
N/A	0	.0051	.0051	0	0	0	0	.0102	.0204
Total	.0918	.0306	.051	.0612	.0306	.1071	.1173	.5102	1

* Number of observation: 196

We did the same analysis for the researchers who were affiliated with the top 10 world-wide universities. According to Table 35, the majority of the respondents had the research budget of more than $300,000. Although the number of publications is relatively high for the high funding group of researchers, their reported number of publications is lower than the NSERC funded researchers. However, we cannot conclude anything since several other factors (*e.g.* quality of the papers, nature of the projects, *etc.*) should be taken into consideration. Again the productivity of the researchers with the lowest level of funding

(<$40K) is not the least where the mentioned reasons can be still valid for the top 10 universities' researchers.

Table 35. Research budget vs. number of publications, top 10 universities' researchers

Research budget	Number of publications								
	<10	10-19	20-29	30-39	40-49	50-69	70-100	>100	Total
< $40K	.0146	.0488	.0244	.0146	.0049	.0049	.0244	.0244	.161
$40K-$80K	.0098	.0098	.0049	.0098	.0049	.0098	.0146	.0195	.0829
$80-$150K	.039	.0146	.0146	0	.0146	.0195	.0049	.0439	.1512
$150K-$300K	.0244	0	.0146	.0195	.0195	.0195	.0341	.0244	.1756
>$300K	.0341	.0341	.039	.0195	.0146	.0537	.039	.0341	.4049
N/A	0	.0049	.0049	0	.0049	0	.0049	0	.0244
Total	.122	.1122	.1024	.0634	.0634	.1073	.122	.3073	1

* Number of observation: 205

In the next part we evaluate the impact of the collaboration motives on the academic and industrial team size of the researchers. For this purpose, we integrated the responses from NSERC and top 10 universities' researchers. Multinomial logistic regression analysis was run on the number of academic and industrial collaborators as the dependent variables. The independent variables were selected based on the motives of collaboration that was rated as the most important factors by the respondents (Figure 69). Various regression analyses were performed where the most robust results are presented in this section.

To analyze the impact of the influencing factors on the academic scientific team size, we considered age and gender as the demographic independent variables. Five most important motives of collaboration (according to the respondents) were also added to the model. As it can be seen in Table 36, age of the researchers has a negative impact on their academic team size. The negative coefficient is larger for the larger academic teams (*i.e.* 26-30, and 30+). This is quite expected since larger teams need more coordination and effort. In other words, it is expected that researchers grow their team size to a certain age and then maintain or decrease the team size after that certain age. Our results suggest a significant positive effect of gender in just one of the categories of academic team sizes (21-25). Hence, in general it seems that gender does not play an important role in selecting academic partners for collaboration. Analysis of the impact of collaboration motives on the academic team size of the researchers reveals that partner's access to special resources and equipment has a significant positive impact on the small academic teams (*i.e.* 6-10, and 11-15). According to the results, interestingly for the large academic teams financial issues and professional

relations are the significant factors that positively affect the number of collaborators in the academic teams.

We did the same analysis to analyze the impact of the demographic and collaboration motives on the industrial team size of the researchers. Number of industrial researchers was considered as the dependent variable. Various variables were included into the model and several logistic regression analyses were performed where the most robust results are presented in Table 37. According to the results, the age variable has a significant impact in two categories of the industrial team members (*i.e.* 21-25, and 26-30) but with different directions. Hence it seems that that there is a limit in the industrial team size of the researchers with respect to their age since large scientific teams (26-30) are negatively influenced by the age of researchers while before the limit (size of 26) it shows a positive impact. This is a preliminary observation that needs further investigation.

The analysis of impact of the collaboration motives reveals that partner's financial ability has a positive impact in small and medium sized industrial teams while it negatively influences large teams. Access to the special resources and equipment shows a negative impact in medium size groups while the coefficient turns to positive for larger industrial teams. On the other hand, past collaboration experience positively affects the medium sized industrial groups while the impact is negative for the larger teams. Hence, it can be said that for the larger teams access to special resources is a more important factor in choosing collaboration partners rather than the past collaboration experience. Interestingly, it seems that relationships do not play a role in forming industrial collaboration teams since a negative coefficient is observed. Lastly, expertise in the complementary field shows a positive impact in large industrial teams. This is quite expected since larger teams might be more interdisciplinary that may include more researchers with various expertises.

Table 36. Impact of demographics and collaboration motives on academic collaboration, logistic regression results

No of academic collaborators	6-10		11-15		16-20		21-25		26-30		30+	
	B	**S.E.**	**B**	**S.E**	**B**	**S.E.**	**B**	**S.E.**	**B**	**S.E.**	**B**	**S.E.**
Demographics												
Age	-.370***	.124	-.45**	.205	-.295	.286	-.202	.447	-3.370***	.868	-.58**	.248
Gender	.388	.308	.809	.587	1.116	1.02	19.411***	.449	-1.43	1.016	.09	.727
Motives												
Partner brings money	-.108	.114	-.472***	.176	-.208	.297	-.541	.397	1.202**	.588	.061	.368
Partner's special resource	.242**	.108	.507***	.138	.047	.245	.172	.527	.14	.41	.284	.248
Personal relations	-.152	.115	-.181	.164	.368	.248	-.644	.401	-.623	.483	.213	.286
Professional relations	-.067	.115	-.04	.163	-.086	.305	-.075	.477	3.505***	1.315	.349	.402
Expertise in complementary field	.096	.112	.082	.170	-.142	.346	.008	.296	.34	.597	.075	.468
_cons	.115	.662	-1.368	1.001	-2.696	1.956	-19.654***	1.82	-16.925***	4.237	-4.371*	2.405

No of observations: 401, Unstandardised Coefficients (*B*), Standard Errors (*S.E.*)

* p<0.10

** p<0.05

*** p<0.01

Table 37. Impact of demographics and collaboration motives on industrial collaboration, logistic regression results

No of industrial collaborators	6-10		11-15		16-20		21-25		26-30		30+	
	B	**S.E.**	**B**	**S.E**	**B**	**S.E.**	**B**	**S.E.**	**B**	**S.E.**	**B**	**S.E.**
Demographics												
Age	.121	.173	.128	.217	.434	.35	6.444***	1.115	-3.987***	.792	.426	.371
Motives												
Partner brings money	.233	.158	.574*	.309	-.1	.544	14.194***	.566	-5.3702***	.9	.825*	.494
Partner's special resource	.193	.138	-.141	.209	-.035	.127	-7.329***	.345	4.229***	.408	.147	.698
Past collaboration experience	-.1	.163	-.137	.399	.415	.328	15.984***	.504	-11.261***	.466	.132	.359
Professional relations	.141	.148	.358	.347	-.674*	.401	-27.344***	.843	-57.335***	1.739	-.688***	.215
Expertise in complementary field	-.102	.165	.246	.497	.178	.491	-17.656***	.603	12.553***	.422	-.228	.563
_cons	-3.117***	.916	-6.606***	2.285	-5.93***	1.596	-55.735***	2.927	-10.859***	3.958	-6.554***	1.2

No of observations: 401, Unstandardised Coefficients (*B*), Standard Errors (*S.E.*)

* p<0.10

** p<0.05

*** p<0.01

5.4.1.5 Conclusion

In this paper we checked for 16 different motives for collaboration. We have found support that there is not a significant difference in collaboration motives among NSERC funded researchers and the researchers' affiliated with the top 10 high ranking world universities. Past collaboration experience and expertise of the potential partner in the complementary field of research were indicated as the two most important motives. Interestingly, the demographic attributes were mentioned as the least influencing factors in selecting collaborators. In addition, the results suggest the importance of the personal and professional relations in finding new partners rather than potential partner's productivity in terms of number and quality of publications.

We also statistically examined the interrelationships between research budget, scientific performance, and collaboration of the researchers. The two-way analysis between research budget and number of publications revealed a positive relation between budget and productivity since the researchers with high amount of research money on average produced more articles. The relation was more significant for the NSERC funded researchers in comparison with the researchers of the top 10 high ranking universities. Moreover, it was observed that researchers with the lowest amount of funding available are not producing the least number of publications. This can be due to the fact that the lowest level of funding group of researchers may maintain or increase their productivity level in order to secure more funding or improve their position.

Analysis of the impact of collaboration motives on scientific team size of the researchers showed that researchers have different motives for forming scientific collaboration of different sizes. Moreover, the impact of collaboration motives in academic and industrial teams is different. In larger academic teams the professional relations and financial motives were the most important factors that affected the size positively. However, smaller academic teams were more influenced by the ability of the potential partner in providing access to the special resources and equipments. Hence, as expected concerns of the researchers for collaborating with others is highly dependent on the size of the team that they are involved in. This is quite reasonable since for example complex projects that are more

interdisciplinary might require a large team of researchers with different expertise that requires more financial investment.

The analysis of the impact of motives on the number of industrial partners of researchers also confirmed the different collaboration concerns of researchers according to their team size. It was observed that researchers with larger industrial partners are more focused on potential partner's expertise in the complementary field and his/her access to special resources when they want to select an industrial partner for collaboration. This is quite expected since industrial firms can provide academic researchers with the special resources while benefitting from academics' expertise as one of the main units in the knowledge diffusion circle. For the smaller teams sizes it was observed that financial resources and past collaboration experience are the two most important motives in finding an industrial partner. Hence it seems that smaller teams prefer to rely on the partners that have a good past collaboration record in order to reduce the collaboration risks.

5.4.1.6 Limitations and Future Work

The first limitation was in regard to the exact numbers of the research budget, publications, collaborators, *etc*. After running the first round of the questionnaires a very low response rate was observed. We held some random interviews with the respondents asking the same questions as the questionnaire and realized that most of the researchers prefer to indicate a range rather than an exact number. Hence, we revised the questionnaire in a way that it contained ranges instead of exact numbers. Although this resulted in a higher response rate, future research can address this issue by asking for the exact numbers.

Moreover, we focused on the NSERC funded researchers and compared their results with the top 10 high ranking world universities. In order to come into a global conclusion it would be informative to focus on the collaboration motives in other countries, institutes, funding organizations, *etc*. It is reasonable to suppose that the collaboration motives and researchers' performance can be influenced by their geographical location and ethnicity. Through this approach and by comparing the results the most important global as well as local motives can be identified.

6. SUMMARY AND CONCLUSIONS

Money is one of the main determinant factors for stimulating research and development activities. Governments are annually investing large amount of money on scientific activities in an aim for improving the socio-economic situation of the country as well as its scientific position world-wide. Limited financial resources in one hand and growing number of researchers on the other hand have made the competition for getting the financial support tighter than ever. Hence, it is needed not only to allocate the available money to the most appropriate and competent applicants but also to evaluate the performance of the funded researchers in respect to the amount of money that they have received.

Apart from the financial resources, scientific collaboration patterns of the researchers can also affect their productivity. Through collaboration researchers can get access to precious external resources (*e.g.* equipments, expertise) that can enhance their overall productivity. However, if collaboration is not managed in a proper way it can harm the scientific output of the team members. Examples can be a member who is not responsible against the deadlines which might affect the performance of the whole team negatively or the coordination costs that can be a serious issue in the large teams.

Hence, to evaluate scientific activities and performance of the researchers it would be more realistic if we consider the inter-relations among funding, scientific collaboration, and the output of the researchers. In this research, the focus was on the mentioned inter-relations at the individual level of the funded researchers for the period of 1996 to 2010. NSERC was selected as the source of funding in this research since it is the main funding organization of the country. The bibliographic data was extracted from Scopus and SCImago was used as the source of annual impact factor of the journals in which the articles were published. A unique data gathering procedure was used in this research to collect and integrate the required data that was explained in the text.

The main purpose of the research was to employ a triangulation technique (using several methodologies) to evaluate the relations comprehensively while proposing a machine learning framework for classification of the researchers as well as predicting their productivity and their deserving level of funding in a given year. For this purpose,

bibliometric indicators and visualization technique were first used to analyze the scientific performance of the funded researchers and their collaboration patterns. In addition, the impact was assessed for different NSERC funding programs and Canadian provinces as well. Moreover, the performance of the researchers affiliated with the top ten Canadian universities was analyzed.

According to the results of the first phase, the Canadian provinces can be divided into two main categories, *i.e.* high funding and low funding groups. The high funding group of provinces contains Ontario, Quebec, British Columbia, and Alberta while the other six provinces belong to the low funding group. The average funding per researcher follows a slightly increasing trend in the low funding group of provinces while in the high funding group, three different periods were observed that were explained in detail in the text. According to the bibliometric results, there was no significant impact of funding on the rate of publications in the high funding group of provinces. However, a positive relation was observed for the low funding group. Interestingly, a positive impact of funding on the quality of the publications was seen for the high funding group of provinces where for the low funding group no relation was observed. Hence, it seems that researchers who reside in the high funding group of provinces focus more on the quality of their works rather than the quantity. One reason can be the higher number of high ranking universities in the former group. In addition, a positive impact of funding on scientific collaboration was observed in the both groups of provinces. Therefore, in general higher level of funding enables researchers to expand their scientific teams.

The analysis of the funding programs revealed that programs that are well-targeted (*e.g.* strategic projects) have resulted not only in higher rate of publications but also higher quality works. This is exactly in line with the definition of such programs since they are mainly allocated to the high-priority research projects that can affect the societal situation of the country. Unlike the other funding programs, these well-targeted programs followed an increasing trend of funding during the whole period of study. Analyzing the relations at the scientific discipline level revealed that researchers in different disciplines have different collaborative behavior. For example, mathematicians prefer to work in smaller teams while health scientists tend to work in large groups.

According to the analysis of the high ranking Canadian universities, the examined universities have almost the same share of publications while the share of funding is lower for the French speaking universities (*i.e.* Université de Monteal, and Université Laval). Interestingly, the level of NSERC funding followed an increasing trend during the whole examined period for all the ten universities. This highlights the important position of the high ranking universities in the R&D system of the country. Finally, a positive impact of funding was observed on the rate and quality of the publications of the researchers as well as their scientific team size.

In the second phase of the research, the interrelations among funding, collaboration, and scientific productivity was statistically analyzed. According to the results, funding, scientific team size, and past productivity of the researchers positively influence their rate of publications as well as the quality of their works. The results suggest the existence of the Matthew Effect in a sense that rich scientists get richer. This finding was also confirmed in the forth paper of the statistical analysis section (*i.e.* funding as the dependent variable). Interestingly, a higher rate of publications was observed for the academic researchers and as expected well-targeted funding programs resulted in higher productivity of the researchers.

It was found that the collaboration network of the NSERC funded researchers strictly exhibits the small world structure. More connected sub-networks, higher number of collaboration links, and easier access to distant information are some of the properties of the small world structure that enables researchers to expand their scientific teams easier. According to the results, funded researchers have benefitted from the small world property to expand their teams and enhance their productivity. The inverted U-shape of the small world trend and the fact that the most recent peak was observed in the period of [2006-2008] highlighted the importance of reevaluating the small world property in the coming years.

Analyzing the impact of the influencing factors on the collaborative behavior of the researchers revealed that academic researchers work in smaller teams in comparison with the non-academics. In addition, a negative impact was observed for the career age of the researchers on their scientific team size that was quite expected. The results suggest that in order to take the gatekeeper role in the collaboration network, researchers should be highly productive in terms of both quantity and quality of the publications. Moreover, higher

amount of money and working in larger teams help researchers to obtain higher betweenness centrality. Since a negative impact was observed for the career age, it can be said that gatekeepers in general are highly productive young or mid-career researchers that have access to financial resources and work in relatively large teams. According to the results for the clustering coefficient model, funding has a negative impact on the formation of triangles and cliques. Hence, researchers may use the financial resources to linearly expand their teams rather than forming highly connected internal communities. As expected, the probability of higher clustering coefficient was higher for the researchers with higher number of connections and no impact of past productivity was observed. Hence to work in a knit group, relations are playing a more important role rather than money or profile of the researchers. The negative impact of funding was also seen for the eigenvector centrality that reflects the leadership role of the researchers. Hence, higher amount of money may involve researchers in other activities (*e.g.* finding right partners to allocate money to) that might harm the leadership role of the researchers. Finally, it was observed that local influencers have high amount of money, are highly productive, work in relatively large teams, and have good relations and links to the other researchers.

Analyzing the effect of the influencing factors on the funding level of researchers suggest that to get higher amount of funding it would be better to be directly connected to a lot of people and work in large teams and tight communities rather than to get connected to highly influential researchers (leaders). In addition, the more important role of the network variables was observed in getting higher amount of funding that partially indicates the determinant role of relations and links in securing higher amount of research money. Being a member of a large highly connected component or locating in Quebec and British Columbia provinces were found as some of the factors that can increase the probability of getting more funding.

Based on the findings of the first and the second phases of the research, the determinant influencing factors were selected and fed into the defined machine learning models to classify the researchers as well as to predict their productivity and funding level. Accuracy of the proposed models was tested with several measures and it was proved that the proposed framework can act as a highly accurate tool in the scientific evaluation procedure. The

proposed tools can help the decision makers to better allocate the in-hand funding and to assess the performance of the researchers more accurately.

Several assumptions were made in this research that were explained in the text. As the complementary phase of the research the assumptions were validated through two-separate survey data analyses. Moreover, the collaboration motives of the researchers were analyzed and compared for the NSERC funded researchers as well as researchers affiliated with the top ten world high ranking universities. The limitations of the research were specifically and separately explained for each of the papers in the text.

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.

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.

Ajiferuke, I., Burell, Q., & Tague, J. (1988). Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics, 14*(5), 421-433.

Albert, R., & Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47.

Albrecht, C. F. (2009). A bibliometric analysis of research publications funded partially by the cancer association of South Africa (CANSA) during a 10-year period (1994-2003). *South African Family Practice, 51*(1).

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One, 4*(6), e5910.

Allison, P. D., & Stewart, J. A. (1974). Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review*, 596-606.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics'. *Journal of Documentation, 53*(4), 404-426.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273-289.

Arnold, E., & Balazs, K. (1998). Methods in the evaluation of publicly funded basic research. *Informe Para La OCDE.*

Arora, A., & Gambardella, A. (1998). The impact of NSF support for basic research in economics. *Economics Working Paper Archive at WUSTL.*

Averch, H. (1990). Policy uses of evaluation of research literature. *OTA Contractor Report.*

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* ACM press New York.

Banal-Estanol, A., Jofre-Bonet, M., & Meissner, C. (2008). The impact of industry collaboration on academic research output: A dynamic panel data analysis. *Working Papers (Universitat Pompeu Fabra.Departamento De Economía y Empresa),* (1190), 1.

Barabâsi, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3), 590-614.

Baum, J. A., Shipilov, A. V., & Rowley, T. J. (2003). Where do small worlds come from? *Industrial and Corporate Change, 12*(4), 697-725.

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.

Beaudry, C., & Clerk-Lamalice, M. (2010). Grants, contracts and networks: What influences biotechnology scientific production? *Danish Research Unit for Industrial Dynamics (DRUID) Conference, London, June,* pp. 16-18.

Beaver, D. 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.

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration. *Scientometrics, 1*(1), 65-84.

Beaver, D. (1984). Teamwork: A step beyond collaboration. *George Sarton Centennial, Communication and Cognition*, 449-452.

Beaver, D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics, 52*(3), 365-377.

Bell, J. G., & Seater, J. J. (1978). Publishing performance: Departmental and individual. *Economic Inquiry, 16*(4), 599-615.

Blume-Kohout, M. E., Kumar, K. B., & Sood, N. (2009). *Federal Life Sciences Funding and University R&D.*

Blumenthal, D., Gluck, M., Louis, K. S., Stoto, M. A., & Wise, D. (1986). University-industry research relationships in biotechnology: Implications for the university. *Science (New York, N.Y.), 232*(4756), 1361-1366.

Bonaccorsi, A., & Daraio, C. (2005). Econometric approaches to the analysis of productivity of R&D systems. *Handbook of quantitative science and technology research* (pp. 51-74) Springer.

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

Bookstein, A. (1980). Explanations of the bibliometric laws. *Collection Management, 3*(2-3), 151-162.

Bordons, M., Gomez, I., Fernandez, M. T., Zulueta, M. A., & Mendez, A. (1996). Local, domestic and international scientific collaboration in biomedical research. *Scientometrics, 37*(2), 279-295.

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

Bossy, M. J. (1995). The last of the litter: Netometrics. *Solaris Information Communication, 2*, 245-250.

Bourke, P., & Martin, B. (1992). Evaluating university research performance-what approach. *What Unit of Analysis.ANU & SPRU.*

Boyack, K. W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology, 54*(5), 447-461.

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

Bradford, S. (1934). Sources of information on specific subjects. engineering, 137, 85-86. reprinted in. *Collection Management, 1*, 95-103.

Braun, D. (2003). Lasting tensions in research policy-making—a delegation problem. *Science and Public Policy, 30*(5), 309-321.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record, 29.* (2) pp. 93-104.

Buss, A. R. (1976). Evaluation of Canadian psychology departments based upon citation and publication counts. *Canadian Psychological Review/Psychologie Canadienne, 17*(2), 143.

Butler, L. (2005). What happens when funding is linked to publication counts? *Handbook of quantitative science and technology research* (pp. 389-405) Springer.

Buxton, M., Hanney, S., & Jones, T. (2004). Estimating the economic value to societies of the impact of health research: A critical review. *Bulletin of the World Health Organization, 82*(10), 733-739.

Campbell, D., & Bertrand, F. (2009). Bibliometrics as a performance measurement tool for the evaluation of research: The case of canadian forest service. *Science-Metrix, 2009 Annual CES Conference.*

Campbell, D., Labrosse, I., Cote, G., & Archambault, E. (2009). Bibliometric assessment of research funded by genome canada 1996-2007. *Science-Metrix.*

Campbell, D., Picard-Aitken, M., Côté, G., Caruso, J., Valentim, R., Edmonds, S., et al. (2010). Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the national cancer institute of Canada. *American Journal of Evaluation, 31*(1), 66-83.

Carayol, N., & Matt, M. (2006). Individual and collective determinants of academic scientists' productivity. *Information Economics and Policy, 18*(1), 55-72.

Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education, 18*(4), 379-389.

Chen, Z., & Guan, J. (2010). The impact of small world on innovation: An empirical study of 16 countries. *Journal of Informetrics, 4*(1), 97-106.

CIHR. (2012). *Canadian institutes of health research.*http://www.cihr-irsc.gc.ca/e/193.html

Clark, B. R., & Clark, B. (1998). *Creating entrepreneurial universities: Organizational pathways of transformation* IAU Press Oxford.

Cole, J. R., & Cole, S. (1973). Social stratification in science.

Cole, S. (1979). Age and scientific performance. *American Journal of Sociology*, 958-977.

Coleman, J. S., & Lazarsfeld, P. F. (1981). *Longitudinal data analysis* Basic Books New York.

Couto, F. M., Grego, T., Pesquita, C., & Verissimo, P. (2009). Handling self-citations using Google Scholar.

Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control, 28*(8), 1557-1575.

Cozzens, S. E., Bobb, K., & Bortagaray, I. (2002). Evaluating the distributional consequences of science and technology policies and programs. *Research Evaluation, 11*(2), 101-107.

Creamer, E. G. (1998). *Assessing faculty publication productivity: Issues of equity. ASHE-ERIC higher education report, volume 26, number 2.* ERIC.

Crespi, G. A., & Geuna, A. (2008). An empirical study of scientific production: A cross country analysis, 1981–2002. *Research Policy, 37*(4), 565-579.

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.

Davis, G. F., Yoo, M., & Baker, W. E. (2003). The small world of the American corporate elite, 1982-2001. *Strategic Organization, 1*(3), 301-326.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics* Scarecrow Press.

De Candolle, A. (1885). *Histoire des sciences et des savants depuis deux siecles.*

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with pajek* Cambridge University Press.

De Solla Price, Derek J. (1965). Is technology historically independent of science? A study in statistical historiography. *Technology and Culture,* , 553-568.

De Solla Price, Derek J, & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist, 21*(11), 1011.

De Solla Price, Derek John, De Solla Price, Derek John, De Solla Price, Derek John, & De Solla Price, Derek John. (1986). *Little science, big science... and beyond* Columbia University Press New York.

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.

Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. *Artificial neural networks and machine Learning–ICANN 2011* (pp. 293-300) Springer.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the Seventh International Conference on Information and Knowledge Management,* pp. 148-155.

Dundar, H., & Lewis, D. R. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39*(6), 607-631.

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

Egghe, L. (2006). An improvement of the H-index: The G-index. *ISSI Newsletter, 2*(1), 8-9.

Egghe, L., & Rousseau, R. (1990). Introduction to informetrics: Quantitative methods in library, documentation and information science.

Egghe, L., & Rousseau, R. (2008). An h-index weighted by citation impact. *Information Processing & Management, 44*(2), 770-780.

Ehrlich, K., & Carboni, I. (2005). Inside social network analysis. *Boston College.*

Eisenhardt, K. M., & Tabrizi, B. N. (1995). Accelerating adaptive processes: Product innovation in the global computer industry. *Administrative Science Quarterly, 40*(1)

Elzinga, A., & Jamison, A. (1995). Changing policy agendas in science and technology. *Handbook of Science and Technology Studies Ed.by Sheila Jasanoff Et Al.(London: Sage).*

Ernø-Kjølhede, E., Husted, K., Mønsted, M., & Wenneberg, S. B. (2001). Managing university research in the triple helix. *Science and Public Policy, 28*(1), 49-55.

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*, 1-21.

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

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal, 22*(8), 2623-2628.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google Scholar: Strengths and weaknesses. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology, 22*(2), 338-342.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 253-290.

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., & Juda, A. I. (2007). Small worlds and regional innovation. *Organization Science, 18*(6), 938-954.

Fleming, L., & Marx, M. (2006). Managing creativity in small worlds. *California Management Review, 48*(4), 6.

Fowler, J. (2005). Turnout in a small world. *Temple University Press, Philadelphia*, 269-287.

Fox, M. F. (1983). Publication productivity among scientists: A critical review. *Social Studies of Science, 13*(2), 285-305.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science* Empirical Press Vancouver.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257-270.

García de Fanelli, A., & Estébanez, M. (2007). Sistema nacional de innovación argentino. estructura, grado de desarrollo y temas pendientes. *Nuevos Documentos Cedes, 31.*

Garfield, E. (1970). Citation indexing for studying science.

Gaughan, M., & Bozeman, B. (2002). Using curriculum vitae to compare some impacts of NSF research grants with research center funding. *Research Evaluation, 11*(1), 17-26.

Gauthier, É. (1998). *Bibliometric analysis of scientific and technological research: A user's guide to the methodology* Science and Technology Redesign Project, Statistics Canada.

Geisler, E. (2004). Measuring the impacts from public sector science and technology: New methods. *Preseantation to the "Workshop on Measuring the Impacts of Science",* Montreal, Canada, June 16-18.

Geuna, A. (2001). The changing rationale for european university research funding: Are there negative unintended consequences? *Journal of Economic Issues*, 607-632.

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*(4), 277-304.

Geuna, A., & Nesta, L. (2003). *University patenting and its effects on academic research,* Citeseer.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies* Sage.

Gingras, Y. (1996). Bibliometric analysis of funded research. A feasibility study. *Report to the Program Evaluation Committee of NSERC.*

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69-115.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69-115.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548-1572.

Godin, B. (1998). Writing Performative History: The New New Atlantis?

Godin, B. (2002). Measuring output: When economics drives science and technology measurements. *Project on the History and Sociology of S&T Statistics, Paper, 14*, 3-27.

Godin, B. (2003). The impact of research grants on the productivity and quality of scientific research. *No. 2003. INRS Working Paper.*

Godin, B., & Doré, C. (2004). Measuring the impacts of science: Beyond the economic dimension. *History and Sociology of S&T Statistics.*

Godin, B., & Gingras, Y. (2000). Impact of collaborative research on academic science. *Science and Public Policy, 27*(1), 65-73.

Gould, R. V., & Roberto, M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology, 19*, 89-126.

Goyal, S., Van Der Leij, Marco J, & Moraga-González, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy, 114*(2), 403-412.

Greene, M. (2007). The demise of the lone author. *Nature, 450*(7173), 1165-1165.

Griffith, R., Redding, S., & Van Reenen, J. (2004). Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. *Review of Economics and Statistics, 86*(4), 883-895.

Gross, P., & Gross, E. (1927). *College libraries and chemical education* Science.

Grossman, J. W. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium,* , 201-212.

Grueber, M., & Studt, T. 2011 global R&D funding forecast: Stability returns to R&D funding. *R&D Magazine (Dec.2010)(Cit.on p.22).*

Guare, J. (1992). *Six degrees of separation* Dramatists Play Service.

Guimerà, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science, 308*(5722), 697-702.

Gulati, R., Sytch, M., & Tatarynowicz, A. (2012). The rise and fall of small worlds: Exploring the dynamics of social structure. *Organization Science, 23*(2), 449-471.

Gulbrandsen, M., & Smeby, J. (2005). Industry funding and university professors' research performance. *Research Policy, 34*(6), 932-950.

Hagedoorn, J., Link, A. N., & Vonortas, N. S. (2000). Research partnerships. *Research Policy, 29*(4), 567-586.

Hall, B. H., Link, A. N., & Scott, J. T. (2001). Barriers inhibiting industry from partnering with universities: Evidence from the advanced technology program. *The Journal of Technology Transfer, 26*(1-2), 87-98.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques,* Morgan Kaufmann.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining* MIT press.

Hanneman, R. A., & Riddle, M. (2011). Concepts and measures for basic network analysis. *The Sage Handbook of Social Network Analysis*, 340-369.

Harriet, Z. (1996). Scientific elite: Nobel laureates in the United States. *New Brunswick: Transaction Publishers.*

Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). *Econometric Models for Count Data with an Application to the Patents-R&D Relationship.*

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.

Heffner, A. G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics, 3*(1), 5-12.

Heinze, T., & Kuhlmann, S. (2008). Across institutional boundaries?: Research collaboration in German public sector nanoscience. *Research Policy, 37*(5), 888-899.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation, 13*(2), 76-86.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569.

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies.*

Hood, W. W., & Wilson, C. S. (1999). The distribution of bibliographic records in databases using different counting methods for duplicate records. *Scientometrics, 46*(3), 473-486.

Horton, R. (1998). Publication and promotion. A fair reward. *Lancet*, 352: 892.

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

Huffman, W. E., & Evenson, R. E. (2005). New econometric evidence on agricultural total factor productivity determinants: Impact of funding composition. *Iowa State University, Department of Economics, Working Paper, 3029.*

Industry Canada. (2002). Achieving excellence: Investing in people, knowledge and opportunity. *Ottawa: Government of Canada, Canada's Innovation Strategy. Industry Canada.Gc.Ca (Accessed on January 22, 2004).*

Ingwersen, P., & Christensen, F. H. (1997). Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues. *Jasis, 48*(3), 205-217.

Irvine, J., & Martin, B. R. (1984). Foresight in science: Picking the winners. *Pinter London*.

ITG. (2008). The science of science policy: A federal research roadmap. *Washington, DC: National Science and Technology Council and the Office of Science and Technology Policy.*

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics, 95*(9), 1168-1177.

Jacob, B., & Lefgren, L. (2007). *The Impact of Research Grant Funding on Scientific Productivity.*

Jin, B. (2006). H-index: An evaluation indicator proposed by scientist. *Science Focus, 1*(1), 8-9.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* Springer.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18.

Katz, R., & Allen, T. J. (1982). Investigating the not invented here (NIH) syndrome: A look at the performance, tenure, and communication patterns of 50 R & D project groups. *R&D Management, 12*(1), 7-20.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science, 13*(5), 261-276.

Kleinman, D. L., & Vallas, S. P. (2001). Science, capitalism, and the rise of the "knowledge worker": The changing structure of knowledge production in the United States. *Theory and Society, 30*(4), 451-492.

Klette, T. J., & Kortum, S. (2002). Innovating firms and aggregate innovation . *No. w8819. National Bureau of Economic Research.*

Kogut, B., & Walker, G. (2001). The small world of Germany and the durability of national networks. *American Sociological Review*, 317-335.

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter, 2*(3), 4-6.

Kostoff, R. N. (2002). Citation analysis of research performer quality. *Scientometrics, 53*(1), 49-71.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning, 30*(2-3), 195-215.

Kumar, S., & Jan, J. M. (2013). Mapping research collaborations in the business and management field in malaysia, 1980–2010. *Scientometrics*, 1-27.

Kyvik, S. (1991). *Productivity in academia: Scientific publishing at Norwegian universities* Rådet for samfunnsvitenskapelig forskning, NAVF.

Kyvik, S., & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics, 76*(3), 439-455.

Landry, R., & Amara, N. (1998). The impact of transaction costs on the institutional structuration of collaborative academic research. *Research Policy, 27*(9), 901-913.

Larivière, V., Gingras, Y., & Archambault, E. (2006). Comparative analysis of networks of collaboration of Canadian researchers in the natural sciences, social sciences and the humanities. *Scientometrics, 68*(3), 519-533.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*(19), 198701.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts* Princeton University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts* Princeton University Press.

Lawani, S. M. (1980). *Quality, Collaboration and Citations in Cancer Research: A Bibliometric Study.*

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly, 52*(4), 667-694.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673-702.

Lehman, H. C. (1953). Age and achievement.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning, 46*(1-3), 423-444.

Leopold, E., May, M., & Paaß, G. (2005). Data mining and text mining for science & technology research. *Handbook of quantitative science and technology research* (pp. 187-213) Springer.

Lewison, G., & Dawson, G. (1998). The effect of funding on the outputs of biomedical research. *Scientometrics, 41*(1), 17-27.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from web-of-science data. *Journal of Informetrics, 6*(2), 318-332.

Leydesdorff, L., & Wagner, C. (2009). Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics, 3*(4), 353-362.

Liefner, I. (2003). Funding, resource allocation, and performance in higher education systems. *Higher Education, 46*(4), 469-489.

Lin, N. (2002). *Social capital: A theory of social structure and action* Cambridge University Press.

Lissoni, F., Llerena, P., & Sanditov, B. (2013). Small worlds in networks of inventors and the role of academics: An analysis of France. *Industry and Innovation, 20*(3), 195-220.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management, 41*(6), 1462-1480.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *BMJ (Clinical Research Ed.), 336*(7645), 655-657.

Lööf, H., & Heshmati, A. (2005). The impact of public funds on private R&D investment: New evidence from a firm level innovation study. *MTT Discussion Papers,* (3).

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences.*

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology, 63*(11), 2140-2145.

Lundberg, J., Tomson, G., Lundkvist, I., Sk? r, J., & Brommels, M. (2006). Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding. *Scientometrics, 69*(3), 575-589.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology & Human Values, 17*(1), 101-126.

Luukkonen-Gronow, T. (1987). Scientific research evaluation: A review of methods and various contexts of their application. *R&D Management, 17*(3), 207-221.

Luwel, M. (2005). The use of input data in the performance analysis of R&D systems. *Handbook of quantitative science and technology research* (pp. 315-338) Springer.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435-444.

Mairesse, J., & Turner, L. (2005). Measurement and Explanation of the Intensity of Co-Publication in Scientific Research: An Analysis at the Laboratory Level.

Mansfield, E. (1995). Academic research underlying industrial innovations: Sources, characteristics, and financing. *The Review of Economics and Statistics*, 55-65.

Martin, B. R. (2003). The changing social contract for science and the evolution of the university. *Science and Innovation: Rethinking the Rationales for Funding and Governance.Edward Elgar, Cheltenham*, 7-29.

Martin, B. R., Salter, A., Hicks, D., & Britain, G. (1996). *The relationship between publicly funded basic research and economic performance: A SPRU review* Science Policy Research Unit, University of Sussex.

Martin, B. R., & Tang, P. (2007). The benefits from publicly funded research. *Science Policy Research Unit, University of Sussex.*

Martin, B., & Etzkowitz, H. (2000). The origin and evolution of the university species. *Organisation of Mode.*

Martín-Sempere, M. J., Rey-Rocha, J., & Garzón-García, B. (2002). The effect of team consolidation on research collaboration and performance of scientists. case study of Spanish university researchers in geology. *Scientometrics, 55*(3), 377-394.

McAllister, P. R., & Narin, F. (1983). Characterization of the research papers of US medical schools. *Journal of the American Society for Information Science, 34*(2), 123-131.

Meadows, A. J. (1974). Communication in science. *Butterworths London*.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy, 29*(1), 31-40.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations* University of Chicago press.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*(1), 60-67.

Mitchell, J. R. (2006). A review of NSERC and SSHRC. *Ottawa, Industry Canada.Available at: Http://www.Ryerson.ca/ors/news_archive/download/Federal_ Review_of%20NSERC _and% 20SSHRC.Doc, Accessed January, 21*, 2009.

Mitchell, T. (1997). Decision tree learning. *Machine Learning, 414*.

Moed, H. F., Van Leeuwen, T. N., & Reedijk, J. (1996). A critical analysis of the journal impact factors of Angewandte chemie and the journal of the American chemical society inaccuracies in published impact factors based on overall citations only. *Scientometrics, 37*(1), 105-116.

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Nalimov, V. V., & Mulchenko, Z. (1969). Naukometriya. izuchenie razvitiya nauki kak informatsionnogo protsessa.[scientometrics. study of the development of science as an information process]. *Moscow: Nauka.(English Translation: 1971.Washington, DC: Foreign Technology Division.US Air Force Systems Command, Wright-Patterson AFB,*

Ohio.(NTIS Report no.AD735-634)) Cited by: Wilson, Conception S.(1999).Informetrics.Annual Review of Informa(TRUNCATED), 34, 107-247.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's co-authorship graph. *ACM Sigmod Record, 32*(3), 8-10.

Nelson, R. R. (2004). The market economy, and the scientific commons. *Research Policy, 33*(3), 455-471.

Newman, M. E. (2000). Models of the small world. *Journal of Statistical Physics, 101*(3-4), 819-841.

Newman, M. E. (2001a). Scientific collaboration networks. I. network construction and fundamental results. *Physical Review E, 64*(1), 016131.

Newman, M. E. (2001b). Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132.

Newman, M. E. (2001c). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.

Newman, M. E. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. *Complex networks* (pp. 337-370) Springer.

Nicolaisen, J. (2002). The J-shaped distribution of citedness. *Journal of Documentation, 58*(4), 383-395.

Noyons, C. (2005). Science maps within a science policy context. *Handbook of quantitative science and technology research* (pp. 237-255) Springer.

NSERC. (2012a). *The natural sciences and engineering research council of canada.*http://www.nserc-crsng.gc.ca/Index_eng.asp

NSERC. (2013). *Report on plans and priorities, 2013-2014.*http://www.nserc-crsng.gc.ca/NSERC-CRSNG/Reports-Rapports/RPP-PPR/2013-2014/index_eng.asp

Okubo, Y. (1997). Bibliometric indicators and analysis of research systems: Methods and examples. *No. 1997/1. OECD Publishing.*

Oyo, B., Williams, D., & Barendsen, E. (2008). A system dynamics tool for higher education funding and quality policy analysis. *Proceedings of the 24th International Conference of the System Dynamics Society.*

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science, 33*(1), 38-43.

Payne, A. A., & Siow, A. (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy, 3*(1).

Peritz, B. C. (1990). The citation impact of funded and unfunded research in economics. *Scientometrics, 19*(3-4), 199-206.

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117-136.

Polster, C. (2007). The nature and implications of the growing importance of research grants to Canadian universities and academics. *Higher Education, 53*(5), 599-622.

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.

Porter, A. L., & Youtie, J. (2009). Where does nanotechnology belong in the map of science? *Nature Nanotechnology, 4*(9), 534-536.

Porter, S. R., & Umbach, P. D. (2001). Analyzing faculty workload data using multilevel modeling. *Research in Higher Education, 42*(2), 171-196.

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.

Prathap, G. (2011). The fractional and harmonic p-indices for multiple authorship. *Scientometrics, 86*(2), 239-244.

Pravdić, N., & Oluić-Vuković, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics, 10*(5), 259-280.

Price, D. d. S. (1963). Big science, little science. *Columbia University, New York*, 119-119.

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292-306.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation, 25*(4), 348-349.

Quinlan, J. R. (1993). *C4. 5: Programs for machine learning*. Morgan Kaufmann.

Rehn, C., & Kronman, U. (2008). Bibliometric handbook for Karolinska Institutet. *Huddinge: Karolinska Institutet.*

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. *JAMA: The Journal of the American Medical Association, 278*(7), 579-585.

Rip, A. (1994). The republic of science in the 1990s. *Higher Education, 28*(1), 3-23.

Roberts, R. E., Flattau, P. E., & Lal, B. (2005). Quantitative models for guiding complex S&T investment strategies. Berlin, Presentation at the International Workshop on the Evaluation of Publicly Funded Research.

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.

Ruegg, R., & Jordan, G. (2007). Overview of evaluation methods for R&D programs. *A Directory of Evaluation Methods Relevant to Technology Development Programs, Prepared for US Department of Energy, Office of Energy Efficiency and Renewable Energy.*

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*(4), 581-603.

Salter, A. J., & Martin, B. R. (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy, 30*(3), 509-532.

Sanz Menéndez, L., & Borrás, S. (2000). Explaining changes and continuity in EU technology policy: The politics of ideas.

Savanur, K., & Srikanth, R. (2010). Modified collaborative coefficient: A new measure for quantifying the degree of research collaboration. *Scientometrics, 84*(2), 365-371.

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.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ (Clinical Research Ed.), 314*(7079), 498-502.

Shapira, P., & Wang, J. (2010). Follow the money. *Nature, 468*(7324), 627-628.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*(11), 758-775.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*(2), 253-280.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology, 41*(1), 643-681.

SSHRC. (2013). *Social sciences and humanities research council.* http://www.sshrc-crsh.gc.ca/about-au_sujet/index-eng.aspx

Statistics Canada. (2010a). *Statistics canada.* http://www.statcan.gc.ca/pub/88-221-x/2010001/t054-eng.htm

Statistics Canada. (2010b). *Statistics canada.* http://www.statcan.gc.ca/pub/88-221-x/2010001/part-partie1-eng.htm

Stephan, P. E. (1996). The economics of science. *Journal of Economic Literature*, 1199-1235.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science, 6*(1), 33-38.

Sullivan, B. N., & Tang, Y. (2012). Small–world networks, absorptive capacity and firm performance: Evidence from the US venture capital industry. *International Journal of Strategic Change Management, 4*(2), 149-175.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management, 28*(1), 1-3.

Tan, D. L. (1986). The assessment of quality in higher education: A critical review of the literature and research. *Research in Higher Education, 24*(3), 223-265.

Tang, L., & Shapira, P. (2012). Effects of international collaboration and knowledge moderation on china's nanotechnology research impacts. *Journal of Technology Management in China, 7*(1), 94-110.

Thorsteinsdóttir, O. H. (2000). External research collaboration in two small science systems. *Scientometrics, 49*(1), 145-160.

Thune, T. (2007). University-industry collaboration: The network embeddedness approach. *Science and Public Policy, 34*(3), 158-168.

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.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry,* , 425-443.

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., Amaral, L. A., & Reed-Tsochas, F. (2007). Small-world networks and management science research: A review. *European Management Review, 4*(2), 77-91.

Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world Problem1. *American Journal of Sociology, 111*(2), 447-504.

Van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & Van Raan, A. F. (2003). The holy grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics, 57*(2), 257-280.

Van Looy, B., Ranga, M., Callaert, J., Debackere, K., & Zimmermann, E. (2004). Combining entrepreneurial and scientific performance in academia: Towards a compounded and reciprocal Matthew-effect? *Research Policy, 33*(3), 425-441.

Van Nierop, E. (2009). Why do statistics journals have low impact factors? *Statistica Neerlandica, 63*(1), 52-62.

Van Raan, A. F. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397-420.

Van Raan, A. F. (1998). The influence of international collaboration on the impact of research results. *Scientometrics, 42*(3), 423-428.

Van Raan, A. F. (1988). Handbook of quantitative studies of science and technology.

Van Raan, A. F. (2005a). Measuring science. *Handbook of quantitative science and technology research* (pp. 19-50) Springer.

Van Raan, A. F. (2005b). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133-143.

Vavakova, B. (1998). The new social contract between governments, universities and society: Has the old one failed? *Minerva, 36*(3), 209-228.

Vinkler, P. (2010). Indicators are the essence of scientometrics and bibliometrics. *Scientometrics, 85*(3), 861-866.

Wang, J., & Shapira, P. (2012). Partnering with universities: A good choice for nanotechnology start-up firms? *Small Business Economics, 38*(2), 197-215.

Wanner, R. A., Lewis, L. S., & Gregorio, D. I. (1981). Research productivity in academia: A comparative study of the sciences, social sciences and humanities. *Sociology of Education,* , 238-253.

Wasserman, S. (1994). *Social network analysis: Methods and applications* Cambridge university press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.

Weiss, S., & Kulikowski, C. (1991). Computer systems that learn.

Wilson, C. S. (1999). Informetrics. *Annual Review of Information Science and Technology (ARIST), 34*, 107-247.

Wray, K. B. (2003). Is science really a young man's game? *Social Studies of Science, 33*(1), 137-149.

Wray, K. B. (2004). An examination of the contributions of young scientists in new fields. *Scientometrics, 61*(1), 117-128.

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

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in china: A coauthorship network analysis. *Scientometrics, 83*(1), 115-131.

Yang, R., & Zhuhadar, L. (2011). Extensions of closeness centrality? *Proceedings of the 49th Annual Southeast Regional Conference,* pp. 304-305.

Yin, R. K. (2009). *Case study research: Design and methods* sage.

Zhang, C. (2009). The e-index, complementing the h-index for excess citations. *PLoS One, 4*(5), e5429.

Zipf, G. K. (1935). The psycho-biology of language.

Zipf, G. K. (1949). Human behavior and the principle of least effort.

Zucker, L. G., & Darby, M. R. (1995). *Virtuous Circles of Productivity: Star Bioscientists and the Institutional Transformation of Industry.*

Zucker, L. G., Darby, M. R., Furner, J., Liu, R. C., & Ma, H. (2007). Minerva unbound: Knowledge stocks, knowledge flows and new knowledge production. *Research Policy, 36*(6), 850-863.

Zuckerman, H. (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review.*

APPENDICES
Appendix A. Published Journal Papers

Journal of Information & Knowledge Management, Vol. 12, No. 4 (2013) 1350037 (16 pages)

DOI: 10.1142/S0219649213500378

World Scientific
www.worldscientific.com

Impact of Funding on Scientific Output and Collaboration. A Survey of Literature

Ashkan Ebadi* and Andrea Schiffauerova†,‡
Concordia Institute for Information Systems Engineering (CIISE)
Concordia University, 1515 Ste-Catherine St. West
Montreal, Quebec, Canada H3G 2W1
*a_ebad@encs.concordia.ca
†andrea@ciise.concordia.ca*

Published 27 December 2013

Abstract. This document critically reviews the papers that investigated the impact of funding on scientific output and on scientific collaboration. For the output, the focus is on the number of articles as a measure of the scientific productivity and the number of citations that a paper received as an indicator of the quality. Various methodological approaches have been adopted (e.g. bibliometrics (a set of methods to analyse the scientific literature quantitatively), statistical analysis) for this purpose. Reviewing the literature revealed that although the general assumption of the positive effect of funding on scientific development is completely (or partially) acknowledged in some studies, one can also find some contradictory results. In addition, we note that analysing the impact of funding on scientific output has attracted more attention of the researchers while investigating the impact of funding on collaboration has been only recently taken into consideration. The paper concludes by comparing the major results and methodologies of the reviewed studies while highlighting the research gaps.

Keywords: Research funding; collaboration; scientific output; literature review.

1. Introduction

Effective funding allocation to highly innovative and original research enables a country to remain at the frontier in the scientific fields, which seems to be crucial for improving its position world-wide in the 21st century. Establishing a well-connected link between the funded research and scientific and knowledge-based systems is very important. Every year, governments spend considerable amounts of money on research mainly through universities and research institutes in order to improve the scientific potential of the country. It is thus essential to analyse the effectiveness of such investment and its impact on the society. In addition, well-defined and effective measures help decision makers to adjust their strategies and to allocate available funding to the most productive and advantageous researchers. Therefore, procedure of evaluating a research needs a group of indicators to create as precise a picture as possible of the various involved aspects in order to assess the performance of a researcher or a research group (King, 1987).

Funding has been acknowledged in many articles to be the main determinant of scientific development (e.g. Martin, 2003) and it is viewed as an important factor that has a significant effect on the scientific output and its quality since it provides a better access to the research resources (Lee and Bozeman, 2005). Several methodologies (e.g. statistical approaches, a variety of indicators, interviews, data and text mining) have been used so far to justify the relation between costs of research and benefits that are gained. Apart from measuring the impact of funding on scientific output, several studies have examined its effect on the rate of scientific collaboration (e.g. Bozeman and Corley, 2004). Moreover, few studies analysed the impact of funding on the patterns of collaboration through creating and forming the networks of the co-operation among researchers (e.g. Defazio *et al.*, 2009). For this purpose, co-authorship analysis has been particularly recognised by some studies (e.g. Glanzel, 2001; Savanur and Srikanth, 2009) as being the most common tool in investigating the co-authorship relations and the quantitative patterns in scientific collaboration.

‡Corresponding author.

This survey sheds light on the key elements that influence the link between allocated grants, scientific output and the structure of scientific collaboration. Although the existing literature mainly confirms a positive effect of funding on scientific activities, there are some gaps in the evidence. This document will highlight some contradictions in the results of the literature. The reasons (e.g. theoretical, methodological) that cause these gaps and defects are addressed, and the areas where further research is required are discussed.

In what follows, first the importance of funding and its major allocation approaches in different countries is discussed. In Sec. 3, the impact of funding on collaboration is discussed and different methodological approaches that have been employed are highlighted. In addition, the respective literature is critically assessed and the main results of the literature are synthesised. In Sec. 4, the literature in regard with analysing the effects of funding on scientific output is critically assessed. Section 5 concludes by identifying main lessons from the reviewed literature and highlighting the research gaps.

2. Funding

Funding and the level of investment have been acknowledged as one of the most crucial factors for improving research productivity (Martin, 2003). More specifically, size, type and sources of funding can be critical factors (Carayol and Matt, 2006). The approach towards the research funding varies across the countries where different procedures are being followed world-wide for funding allocation. This also proves the importance of the mechanism of funding for scientific and technological improvement (Rosenberg, 1976; Jaffe, 1989; Mansfield, 1995; Zucker *et al.*, 1998). For this purpose, performance based (e.g. in UK) or educational size based approaches or a combination of the two systems (e.g. in The Netherlands, Finland and Denmark) have been adopted in different countries.

Both methods have advantages and drawbacks. Performance-based evaluations can enhance efficiency in short-run and create a better accountability. Moreover, they can be used for relating research to policy (Martin and Geuna, 2003). However, this evaluation method has some disadvantages. Getting reliable information to perform the evaluation is highly expensive (Bourke and Martin, 1992). In addition, if one can earn more from research rather than from teaching by a performance-based funding system, professors will tend to the former (Martin and Geuna, 2003), which may influence the training procedure of skilful graduates in academic environments.

To overcome the mentioned limitations some countries (organisations) tend to adopt educational size based funding systems, also popular due to the simplicity in application and low cost of operation. However, they have also some drawbacks. These systems can give a very high power to the distributors of funds. In addition, it is hard to relate the number of the students to the scientific effort of that department (Martin and Geuna, 2003).

In the following section, the scientific collaboration is first defined. Then, the literature studying the impact of funding on the scientific collaboration is examined, compared and discussed.

3. Funding Impact on Scientific Collaboration

Scientific activities know no borders. All researchers worldwide are working together in a global community to improve the level of knowledge. Technological developments are the applications of scientific knowledge where the scientific activities highly rely on the prior knowledge (Subramanyam, 1983). Due to the nature of the modern science which has become more complex and inter-disciplinary, scientists may tend to collaborate more.

According to Katz and Martin (1997), scientific collaboration is defined as the process through which the researchers with a common goal work together to produce new scientific knowledge. Scientific collaboration has been studied in a vast number of different disciplines such as computer science, sociology, research policy and philosophy (Sonnenwald, 2008). In addition, various types of collaboration have been mentioned in the literature including inter-firm collaboration, international collaboration and academic collaboration (Subramanyam, 1983). Due to the diversity in examining the scientific collaboration and its different types, various methods, approaches and terminologies can be found for this purpose in the literature (Sonnenwald, 2008).

Measuring the scientific collaboration is not easy. Although co-authorship and sub-authorship[1] have both been considered in the literature as indicators of scientific collaboration, only 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 and Maffioli, 2011). Co-authorship as a measure is practical, invariant, verifiable, inexpensive (Subramanyam, 1983) and quantifiable (Katz and Martin, 1997). Through co-authorship researchers get access to an

[1] Measured by the number of researchers/colleagues thanked in the acknowledgement section (Sonnenwald, 2008).

often informal network of scientists that may facilitate knowledge and skill diffusion (Tijssen *et al.*, 1996; Tijssen, 2004; Calero *et al.*, 2005). There are also some drawbacks in using co-authorship as an indicator of collaboration. One of them is that collaboration does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz and Martin, 1997).

Collaboration can generate large advantages for the society. Through the collaborative scientific activities, different skills and ideas are combined, and resources are thus used more efficiently. This can bring economies of scale in scientific activities and may avoid research duplication (Ubfal and Maffioli, 2011). In addition, collaboration trains the available skills that will result in the development of new expertise (Lee and Bozeman, 2005). However, there are some costs (e.g. finding right partners and research coordination) related to the scientific collaboration that may influence the optimal individual collaboration level (He *et al.*, 2009). Cummings and Kiesler (2007) focused on the effects of the coordination costs on collaboration among US universities and found that coordination failures have a negative impact on scientific collaboration. However, Adams *et al.* (2005) evidenced that the scientific collaboration cost has declined in the last two decades, which might be explained by the lower travelling costs and improved communication technology.

Although governmental funding for knowledge creation and diffusion has a long history, its effects on scientific collaboration and formation of scientific networks is relatively new (Katz and Martin, 1997; Lee and Bozeman, 2005). The importance of collaborative research is now acknowledged in scientific communities (Wray, 2006), where financial investment can change the structure of research groups and affect the collaboration among the scientists. However, there might be some conflicts between individual preferences and the society level goals. These conflicts may cause different optimal individual collaboration level from the optimal social one. Therefore, the efficient collaboration network will not be stable since the central actor(s)[2] bears a huge coordination cost that is not of his/her private interest. As a result, it is important to evaluate the policies that affect the collaboration, the relation between the individual incentives and social benefits (Ubfal and Maffioli, 2011).

A great advantage of public funding is that it enables researchers to cover the collaboration costs. Moreover, it allows the central actors to better internalise some of the required duties through the coordination (Ubfal and Maffioli, 2011). According to Porac *et al.* (2004), availability of funding may help the central scientist(s) to make a balance between the new knowledge creation and the management of the existing collaborative relationships in the network. On the other hand, one may notice that higher amount of funding is not always beneficial. If the collaboration network is at its social optimum level then allocating more funding can affect the system negatively by adding more collaboration links (Ubfal and Maffioli, 2011). This scenario normally happens in developed countries and is not the case in developing countries (García de Fanelli and Estébanez, 2007).

There are only a few studies that specifically investigated the effects of funding on scientific collaboration. However, most of them are limited to the performance of universities or educational environments. Also, they have not considered a test group of non-funded researchers. This can help to determine the net impact of funding. In addition, to our knowledge no efforts have been made towards analysing the funding impact on the structure and pattern of collaboration networks.

In an early study, Beaver and Rosen (1979) studied the effect of funding on the average number of authors per article as an indicator of the scientific collaboration in 24 scientific fields, and found a positive relation between funding and the average number of authors per article. Two years later, Heffner (1981) collected 500 articles published in 28 journals in four scientific fields during 1974–1975 and analysed the relation between funding and multiple authorship statistically. He found that with an increase in the financial support the size of the research teams has generally increased, but the impact of funding was statistically significant just in chemistry and biology (two fields out of the four examined fields).

Using questionnaires for gathering data and performing regression analysis, Bozeman and Corley (2004) analysed the collaboration among a group of scientists affiliated with universities in the US and found a significant positive effect of funding on their collaboration. Using different independent variables, Adams *et al.* (2005) performed another analysis and found that the researchers of top US universities that have larger amounts of federal funding available tend to work in larger scientific groups, which confirms the findings of Bozeman and Corley (2004). Gulbrandsen and Smeby (2005) considered similar independent variables as Bozeman and Corley (2004) and Adams *et al.* (2005) to study the effect of industry funding on the performance of university professors in Norway.

[2]More central actors have more connections and tend to have favored positions in the network.

However, they applied a qualitative approach and used questionnaires to collect data from all tenured professors in Norway. In addition, they employed logistic regression rather than Ordinary Least Squares (OLS) linear regression. They observed that funded professors tend to collaborate more with other researchers from both universities and industries, which confirmed the finding of the mentioned previous studies.

In another attempt on investigating the university–industry link, Lundberg *et al.* (2006) analysed the effectiveness of co-authorship analysis in measuring university–industry collaboration and the impact of industrial funding on such collaboration. They focused on the industrial collaboration at Karolinska Institutet (KI), located in Stockholm (Sweden). Using indicators they compared the co-authorship data of KI with the industrial funding that was allocated to it. Their analysis includes 436 industrial companies that provided funding to the university. They found that two thirds of the companies had co-authored at least one publication with the university. They also tried to confirm their findings from the companies' side and found that just 16% of the companies had co-authored publications. They concluded that their results are incomplete since they realised a conflict between the funding and co-authorship indicators.

Thune (2007) qualitatively analysed the micro-dynamics of the collaboration among universities and industry using a social capital perspective. He concluded that social capital resources are important in forming collaborative projects. However, forming a successful collaboration is very difficult since it depends on a vast variety of other factors such as trust, familiarity, common language and understanding. In another study and employing logistic regression, Rosenzweig *et al.* (2008) analysed the American papers published in the Academic Emergency Medicine, Annals of Emergency Medicine, Journal of Emergency Medicine and American Journal of Emergency between 1994 and 2003. Although the collaboration as indicated by number of authors per paper increased during the examined period, they found no significant relation between collaboration and extramural funding.

Recently, Defazio *et al.* (2009) studied the impacts of funding on collaborative behaviour and productivity of the researchers in the European Union (EU) funding program framework considering different funding periods in their analysis. In a 15-year period, they used a panel of 294 scientists in 39 EU research networks. They concluded that public funding may play an important role in forming more effective collaboration networks in EU region. The summary of the respective papers are depicted in Table 1.

As it is shown in Table 1, researchers have started evaluating the impact of funding on the collaboration using simple indicators in the early 1980s. Although their results mostly show a positive impact of funding, the approach that was used is not sufficiently rigorous to make any conclusion since they were mainly based on very simple indicators. In addition, their datasets were very limited. After a considerable time gap, researchers have started using more complex and integrated methods for analysing the effect. One of the reasons could be the availability of the data thanks to the progress in information technology. However, still some limitations in the methodologies and datasets can be seen. Lewison and Dawson (1998) did a statistical analysis on biomedical papers. One of the main limitations of their study is that they have considered the journal impact factor as their paper quality proxy instead of the number of citations which is a more common practice in the literature. Using journal impact factor has several drawbacks (e.g. it is highly discipline dependent, editorial policies may affect the impact factor) and it is not accepted as a good paper quality measure (Moed and van Leeuwen, 1996; Seglen, 1997). Bozeman and Corley (2004) used questionnaire (response rate of 45%) to gather their data and then did OLS regression on the collected data. The main concern about their study is the data since it is not representing all the university scientists and is very limited. This limitation can be also seen in Gulbrandsen and Smeby (2005).

Although Adams *et al.* (2005) studied an enormous collection of papers, they have not considered a time window to specifically analyse the effect of being funded in the current time window on the number of papers in the next year(s). In addition, the impact of funding has not been investigated at the individual level and it is limited to the university-field. Among all the studies mentioned, only Defazio *et al.* (2009) defined three different variables for funding reflecting pre-funding, during funding and post-funding effects. However, they just focused on some well-known funded researchers.

Despite some studies that found a positive effect of funding, Lundberg *et al.* (2006) and Rosenzweig *et al.* (2008) could not find any significant relation between funding and collaboration. The reason may be that they have used a limited data source for their analysis. As an example, Rosenzweig *et al.* (2008) just considered the papers of four general peer-reviewed journals published in the US. Thune (2007) used a qualitative approach for addressing the problem and using data from interviewing 29 researchers found a vague effect of funding.

As mentioned above, most of the studies employed statistical analysis. Table 2 shows important issues that

Table 1. Summary of the research on evaluating the impact of funding on scientific collaboration.

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
Beaver and Rosen	1979	Indicators	—	24 scientific fields	Positive relation
Heffner	1981	Statistics	• 395 articles • 28 journals • 1974–1975	4 scientific fields	Positive relation found in two (out of four) disciplines
Bozeman and Corley	2004	• Questionnaire • Regression analysis	451 scientists and engineers in the US	Scientists' collaboration choices	Significantly positive
Adams *et al.*	2005	Regression analysis	2.4 million papers 1981–1999	Top 110 US universities	Positive effect of funding on team size
Gulbrandsen and Smeby	2005	• Questionnaires • Logistic regression analysis	1,967 records	Tenured university professors in Norway	A positive relation observed
Lundberg *et al.*	2006	Indicators	• Industrial funding to a medical university • 1993–2003	Co-authorship between university and industry	• Incomplete results • Some signs of positive effect on collaboration
Thune	2007	Qualitative analysis	Interviews with 29 researchers and R&D managers	Collaborative R&D projects in two academic fields	Important, but other factors are also involved
Rosenzweig *et al.*	2008	Logistic regression	5,728 articles published in 4 American journals 1994–2003	US Emergency Medicine (EM)	No significant relation
Defazio *et al.*	2009	Regression analysis	• Panel of 294 scientists • 39 EU research networks	Researchers in the EU funding program framework	Funding may affect the formation of more effective collaboration networks

Table 2. Statistical analyses and variables, impact of funding on scientific collaboration.

Article	Ctrl. grp.	Independent variables*						Type of analysis	Sources of funding	Different funding periods	Network structure
		Fund	Gender	Prestige/ career/age	Scientific fields	Regional share	Past productivity				
Heffner (1981)	✓	✓						Descriptive analysis	No	No	No
Bozeman and Corley (2004)		✓	✓	✓	✓			OLS Linear regression	No	No	No
Adams *et al.* (2005)					✓	✓		Linear regression	No	No	No
Gulbrandsen and Smeby (2005)		✓	✓	✓	✓	✓		Logistic regression	Yes	No	No
Rosenzweig *et al.* (2008)					✓			Logistic regression	Yes	No	No
Defazio *et al.* (2009)						✓	✓	Linear regression	No	Yes	No

*No. of co-authors (or similar variables such authors/paper) have been considered as the dependent variable.

have been considered as the crucial determinant factors in the studies that applied statistical analysis. As it can be seen, regional share and scientific fields variables have attracted more attention of the researchers. Gulbrandsen and Smeby (2005) was the most comprehensive study done from the number of independent variables and methodology points of view.

To be able to calculate the net impact of funding on collaboration several factors are required to be taken into consideration, e.g. various sources of funding data, a control group of non-funded researchers, or different funding periods. If one neglects this kind of variables in the analysis then it would be hard to conclude that the resulting impact is directly associated with the level of funding. According to Table 2, among all the reviewed studies Gulbrandsen and Smeby (2005) considered a control group of researchers who applied for the funding but were not selected (Heffner (1981) has also used a control group but did not perform any regressions), but they have not included different funding periods. Therefore, the net impact of funding on collaboration is still vague and calls for more analyses. For this purpose, availability of considerable digital datasets can help researchers to analyse the effect more comprehensively at the individual level. In the next section, the literature that has studied the impact of scientific collaboration on the output of the researchers is surveyed and discussed.

4. Funding Impact on Scientific Output

Investigating the impact of funding on the quality and quantity of the published research has attracted more attention of the scientometrists in comparison with analysing funding effect on collaboration. It is easy to judge the productivity and the impact of the research of the Nobel laureates. However, for the rest of scientists one should have quantitative indicators in order to analyse and compare the scientific productivity of the researchers (Hirsch, 2005). The number of publications has been widely used in the literature as the quantity proxy of scientific activities. However, it has been suggested (Okubo, 1997) that publication counts can be considered as a reliable measure just for large-scale data (e.g. macro-level, cross-countries level).

It is generally accepted in bibliometrics that the real or expected number of citations received by publications can be used as a good index of the mean impact at the aggregate level (Gingras, 1996). However, the citations have several drawbacks and thus are considered by some (e.g. Seglen, 1992) as a poor measure of quality. As an example, papers of famous scientists are more likely to be cited. One of the reasons is that they normally supervise a lot of students and they have different teams working on various projects. Apart from that, a low quality work may receive many negative citations, i.e. it is cited not due to its quality but due to an error in methodology or results (Okubo, 1997).

Evaluating the relation between the research input (e.g. research funding) and the research output (e.g. number of publications) has been a challenging issue for policy makers. A number of techniques (e.g. scientometrics, statistical analysis) can be used for this evaluation (King, 1987). It is generally agreed that funding has a positive effect on scientific development and number of scientific publications (McAllister and Narin, 1983; Boyack and Borner, 2002; Godin, 2003; Campbell *et al.*, 2009, 2010). Apart from the number of publications, the impact of funding on the quality of published papers has been also studied. In this section, first we discuss the studies that analysed the impact of funding on the quantity and quality of the output. Then, the studies that have evaluated the impact of funding at macro-level are investigated and, finally, we will specifically discuss the studies that have been done in Canada so far.

4.1. *Funding impact on quantity and quality of scientific output*

In an early case study performed by McAllister and Narin (1983) for the National Institute of Health (NIH) the relation between NIH's funding and number of publications of the US medical schools was investigated. Using bibliometric indicators they found a quite strong relationship between the funding and the number of papers published. Moreover, they found that the number of papers and their citations for each medical school are well related to the quality of the school. Their results partially indicated that funded research may be more cited than the unfunded one. In a similar study Peritz (1990) analysed the citation impact of funded and unfunded research in the field of economics using statistical analysis. He found that even if both funded and unfunded research results are published in a high-impact journal, the funded research will be more cited, which is in line with the findings of McAllister and Narin (1983). Although he found a positive impact of funding on the quality of the output, the approach he used has been criticised in the literature (e.g. Freedman *et al.*, 1978). Peritz (1990) performed a significance test but he did not consider a random model, in which case the significance test may overstate the accuracy of the results (Freedman *et al.*, 1978).

A few studies were focused on the effect of funding on the output of medical (health) schools (programs). Lewison

and Dawson (1998) studied the funding effects on the outputs of biomedical research. They used journal impact factors as a quality measure with a small modification (applied a five-year citation period) to overcome the short-term influence problem which such indicators may have. They concluded that the number of authors per article and the number of funding bodies both have a great effect on the impact of research output. With the increase in the number of authors, an increase in multi-disciplinarity could be observed. This is considered to be a highly important factor for an increase in the impact of the research output. More specifically, they found that if the number of authors rises from one to six then the mean journal impact increases more than twice while the number of citations received is tripled. Jacob and Lefgren (2007) analysed the effectiveness of government expenditures in R&D by investigating the impact of NIH funding on the quantity and quality of the papers of the funded researchers. Their database contained researchers who were funded by NIH in 1980–2000. They used OLS regression to perform the analysis. According to their results, NIH grants had a positive impact on the publication rate leading to about one additional publication over the next five years. This positive impact was higher for postdoctoral fellows. Cancer Association of South Africa (CANSA) conducted a research to evaluate the quality and quantity of the funded research during a 10-year period (1994–2003). In this study, Albrecht (2009) took advantage of bibliometrics and counted the number of peer-reviewed publications in PubMed database for each grantee which were also related to CANSA grants. Since CANSA grants were partial he could not create a benchmark for the cost of an average, peer-reviewed cancer research publication in South Africa. However, he found that the research was more focused on the areas of cancer biology and experimental treatment.

Arora *et al.* (2000) analysed the impact of the contractual funding on Italian academic researchers who work in the biotechnology field. They defined a scientific research production function including various variables such as budget requested, budget granted, size of the group, age of the principal investigator (PI) and number of papers adjusted by quality. They found that although the average elasticity[3] of the output with respect to the funding is around 0.6, the most reputed research groups have elasticity close to 1. In addition, they realised more unequal distribution of funds may increase the output in the short term. However, they had some limitations in performing their analysis such as lack of micro-level data on funding levels and research output in various scientific fields. Carayol and Matt (2006) studied some important factors that affect quantity and quality of scientific production of the faculty members of Louis Pasteur University. Based on their funding variables, they concluded that the effect of private contractual funding is not significant. However, research output is positively influenced by the public contractual funding. But even in this case the respective coefficient is very small.

Payne and Siow (2003) analysed the impact of federal funding on 74 research universities. Employing a regression analysis on a panel dataset spanning from 1972 to 1998, they investigated the effects of funding on the articles published and patents issued by the researchers. Their results show a small positive impact of funding on the number of patents while the effect on the number of articles is relatively higher ($1 million leads to 11 more articles and 0.2 more patents). They could not find a significant impact of funding on the quality of the articles measured by number of citations per article. In an econometric evaluation of the impact of funding composition on agricultural productivity, Huffman and Evenson (2005) used annual data for 48 US states from 1970 to 1999. They found a significant negative impact of the federal competitive grant funding on the productivity of public agricultural researchers. Gulbrandsen and Smeby (2005) studied the effect of industry funding on the performance of university professors in Norway. They used questionnaires to collect data from all tenured professors in Norway and employed logistic regression to perform the analysis and found a positive relation between the industry funding and researchers' performance.

In two recent studies, Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) studied the impact of public and private funding on the scientific production of the Canadian academics working in biotechnology and nanotechnology fields respectively. Beaudry and Clerk-Lamalice (2010) defined their regression model including structural network properties variables, universities dummy variables, and grant and contract amounts. They found no negative impact of contracts on the publication output of researchers working in biotechnology. However, a positive effect of funding and strong network position on the scientific output was observed. The regression model of Beaudry and Allaoui (2012) is very similar to the Beaudry and Clerk-Lamalice (2010) model. They added number of patents and age of the researchers variables to their model and assessed the impact of the considered variables on the scientific output of nanotechnology researchers. Although they found a

[3] Elasticity measures how changing one variable affect other variables. Small changes can have large effects on elastic variables.

positive effect of public funding on scientific publications, the effect of private funding on scientific output is non-existent.

In a quite distinct attempt to relate funding to the scientific development, researchers have recently focused on 3D mapping of the grants and publication data using the visualisation tools such as VxInsight©. Boyack and Borner (2002) evaluated the influence of grants on publications. By using VxInsight map, they found a positive relation between the allocated funds and the publication rate in most of the cases. They have also included a 3D map combining the grant and publication data both in one picture trying to better visualise the impacts. They propose that although such resulting maps cannot replace human decision making, the researcher or government workers can use them to accelerate their understanding of large datasets and to facilitate the decision making procedure. This is a pioneer study in 3D visualising of funding and scientific output data together in one map. However, they faced with some limitations such as lack of more accurate data and using larger amounts of data that might need more efficient and more complex data mining and clustering techniques. In the next section, the literature that analysed the impact of funding on scientific output at the cross-countries level (macro-level) is reviewed.

4.2. *Funding impact on scientific output, macro-level*

Some researchers have studied the impact of financial investment on scientific production at cross-country level. Leydesdorff and Wagner (2009) analysed the relation between research macro-level investment and world share of publication. They employed the Main S&T Indicators of OECD (2008). They found a lot of differences among examined countries in terms of their efficiency in turning financial investment into scientific output. Apart from the efficiency issue, they found different schemes of funding in various countries. In an econometric study, Crespi and Geuna (2008) analysed the important factors (especially the investment) that influence scientific productivity. They focused on the higher education in 14 OECD countries and used Thomson ISI database to gather the publication and citation data for the period of 1981–2002. They mainly focused on the time lag structure of the output and the nature of the spillovers and concluded that investment had a significant impact. In a very brief study, Shapira and Wang (2010) investigated the impact of nanotechnology funding. They used Thomson Reuters database for the period of August 2008 to July 2009 and used very basic bibliometric indicators to give a general picture of countries which are working in nanotechnology field. They argued that as an impact of large investment that has been made, China is getting closer to the US in terms of the number of publications but Chinese papers still have lower quality in comparison with the Americans and Europeans.

4.3. *Funding impact on scientific output, case of Canada*

The evaluation of research performance in Canada has started attracting the attention of the policy makers recently. In Canada, scientific articles have been recognised as the main output of researchers and universities (Godin, 2003) and bibliometrics has been used for scientific evaluation purposes. This section discusses the studies that investigated the role of funding and investment on the productivity of the researchers in Canada so far. Gingras (1996) in a report to the Program Evaluation Committee of NSERC discussed the feasibility of bibliometric evaluation of the funded research. He focused on two grant selection committees (the Mechanical Engineering and Evolution & Ecology) aiming to find whether the results of the indicators and data which come from bibliometrics can be used for answering the questions about funding allocation policies. He showed that it is applicable to use bibliometric indicators to investigate the relation between funds and scientific productivity since these measures give considerable information about such relations. Furthermore, his analysis indicated that it is feasible to apply evaluative bibliometric methods on the funded research at the disciplines or specialties level. Gingras (1996) employed simple bibliometric indicators for performing his analysis and did not perform any statistical analysis.

Following his study, a few Canadian researchers used bibliometrics for analysing the funding impact. Godin (2003) in a bibliometric evaluation studied the impact of NSERC funding on the productivity and papers' quality of the supported researchers for the period of 1990–1999. He used Science Citation Index (SCI) database and analysed the number of papers written by funded researchers over a 10-year time period to find NSERC proportion amount of contribution to the scientific development of Canada. For this purpose, he applied two indicators (the ratio of papers of funded researchers which were written in collaboration with others and the journal quality in which the funded researchers publish their papers). He found that researchers with higher amount of funding available are more productive. In addition, when the level of funding for a given researcher is above the median (high) his/her productivity is more strongly correlated with the

amount of funding. However, the level of funding does not affect the researchers' journals quality. These results are based on simple bibliometric indicators, hence no strategy in regard with better allocation and distribution of grants can be set. In a series of case studies, Campbell *et al.* performed bibliometric evaluations of the impact of funding on scientific performance. Campbell *et al.* (2010) utilised bibliometrics as the performance measurement tool for evaluating the impacts of the research which was funded by the National Cancer Institute of Canada (NCIC). They worked on Thomson Reuters' Web of Science (WoS) database (which includes three databases: SCI Expanded, the Social Sciences Citation Index and the Art & Humanities Citation Index) trying to cover all fields of science to get the respective statistics of NCIC funded researchers. They calculated two main bibliometric indicators in this respect, i.e. number of papers and average of relative citation (ARC). Besides, some statistical tests were performed to check the differences between the scientific impacts of different entries. Their findings show a positive relation between the funds that have been provided by NCIC and the scientific performance. In other words, the ARC of NCIC-supported papers were of higher value than those of non-supported ones. In a conference presentation Campbell and Bertrand (2009) reviewed the results of a bibliometric measurement of research performance for the Canadian Forest Service (CFS). They used a quite wide range of bibliometric indicators to assess CFS internal, national and international position. They found that CFS has the most papers published in forestry in Canada and internationally it ranks as 3rd. Although there were some fluctuations in the impact of CFS publications during 1990–2002, it has increased to above the world level during 2003–2006. Finally, they found that CFS ranked 3rd in number of collaborations within the top 50 world institutions network. In another similar research, Campbell *et al.* (2009) evaluated specifically the selection procedure of Genome Canada to see whether it allocates the funds to the right researchers. By means of Thomson Reuters' WoS database and many bibliometric indicators (number of publications, ARC, ARIF, number of cited papers, number of papers in highest impact journals), they claimed that the peer-review process of Genome Canada was successful in researchers selection. Moreover, the papers published by Genome-funded researchers have a significantly higher scientific impact than other genomics papers not just in Canada, but also all over the world. That means there was a positive relation between Genome-funded projects and the scientific performance of the supported researchers. Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) who also studied the impact of funding in Canada, were discussed before in Sec. 4.1.

4.4. *Funding impact on scientific output, summary*

For over sixty years, governments have funded researches (Godin and Doré, 2004) and have used various tools and techniques, quantitative and qualitative, to measure the scientific performance (Godin, 2002). This section summarises the reviewed literature about the subject.

As it can be seen in Table 3, the most dominant approach that has been used in analysing the impact of funding on the scientific productivity is the statistical analysis. Table 4 shows different determinant factors that have been considered in the literature in the respective statistical models. Similar to what we have seen before in Table 2, regional share and scientific field variables have been the most attractive variables for the researchers while paper quality has been also considered as an important factor. From Table 3 and by counting the number of recent studies, it is clear that this issue is becoming more important and is getting more attention of the researchers. Some reasons could be the increase in the number of the authors, the limited sources of funding available and recent economic depressions that forced the policy makers to re-evaluate their strategies in order to motivate the scientific development more efficiently.

Among the studies that performed regression analysis Arora *et al.* (2000) used a control group of non-funded researchers. However, they have not used it to assess the net impact of funding on the scientific output, but they employed the information from non-funded units to estimate the grant selection and resource allocation equation. Since the other two studies that used the control group (Peritz, 1990; Campbell *et al.*, 2010) performed descriptive analysis, the net impact of funding on scientific output is still vague. In other words, apart from the funding variables that directly affect the scientific productivity, other factors can also affect the output indirectly. For example, higher funding may influence scientific collaboration by forming more efficient groups that may lead to higher scientific productivity. Therefore, one may notice that the collaboration network structural variables can be important factors in determining the productivity of the researchers. Beaudry and Clerk-Lamalice (2010) and Beaudry and Allaoui (2012) have recently considered this issue in their study and added two network structure variables.

Table 3. Summary of research on evaluating the impact of funding on scientific output.

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
McAllister and Narin	1983	Bibliometric indicators	—	NIH funded researchers	• Strong relationship • High quality schools are more productive
Peritz	1990	Statistical analysis	Articles published in two British journals in 1978 and 1979 and their citations	Economics	Funded researchers are being more cited
Gingras	1996	• Feasibility study • Bibliometric indicators	1984–1993	NSERC funded research	Feasible to apply evaluative bibliometric methods on the funded research at the disciplines or specialities level
Lewison and Dawson	1998	Statistical analysis	Research Outputs Database (ROD) 12,925 UK papers 1988–1994	Biomedical (gastroenterology)	Positive relation between the number of funding bodies and research output impact
Arora *et al.*	2000	• Econometrics • OLS regression	1989–1993	Biotechnology, bioinstrumentation Researchers sponsored by the Italian National Research Council (CNR)	More unequal distribution of funds may increase the output in the short term
Boyack and Borner	2002	Visualisation 3D mapping	33,448 grants records 4,549 outputs 1975–2001	Behavioural and Social Science (BSR) program Output resulted from National Institute on Aging (NIA) grants	Positive relation between funding and output in most of their cases
Godin	2003	Bibliometrics	SCI database 1990–1999	NSERC funded research	Positive relation between funding and productivity, but no impact on quality
Payne and Siow	2003	Regression analysis	Federal funding 1972–1998	74 research universities	• Small positive effect on the number of patent and a larger positive effect on the number of articles • No impact on research quality
Huffman and Evenson	2005	Econometrics	USDA 1970–1999	Agricultural research productivity	Negative impact of the federal competitive grant funding on the productivity of public agricultural researchers
Carayol and Matt	2006	• Statistical analysis • Linear regression • OLS model	1993–2000	Faculty members of Louis Pasteur University (ULP)	No significant effect of contractual funding, except for the public one where the coefficient is a small positive number
Jacob and Lefgren	2007	OLS and regression	1980–2000	NIH funded researchers	Positive impact on the rate of publication

1350037-10

Table 3. (*Continued*)

Author(s)	Year	Type of analysis	Data	Target area	Result(s)
Crespi and Geuna	2008	Econometric	Thomson ISI database 1981–2002	Higher education in 14 OECD countries	Investments have a significant impact on the time lag structure of the output
Albrecht	2009	Bibliometric indicators	PubMed database 1994–2003	CANSA	No conclusion in regard with impact of funding on output
Leydesdorff and Wagner	2009	Main S&T indicators of OECD (2008)	Thomson's WoS	Macro-level comparisons	At cross-country level, lots of differences are observed in the link between investment and world-share of publication
Campbell *et al.*	2010	Bibliometrics	Thomson Reuters' WoS	Researchers funded by NCIC	Positive relation between funds and output quality
Shapira and Wang	2010	Bibliometric indicators	Thomson ISI database 2008–2009	Cross-country evaluation	Positive impact of China's investment on output quantity, but no major effect on the quality
Beaudry and Clerk-Lamalice	2010	Regression analysis	1985–2005 3,678 articles	Canadian biotechnology academics	• No negative impact of contracts on publication. • Positive effect of strong network position and individual funding on scientific output
Beaudry and Allaoui	2012	Regression analysis	Articles and patents 1985–2005: • 3,724 articles • 566 patents	Canadian nanotechnology researchers	Positive effect of public funding, no impact if private funding

Table 4. Statistical analyses and variables, impact of funding on scientific output.

Article	Ctrl. grp.	Independent variables*							Type of analysis	Various sources of funding	Different funding periods	Network structure
		Fund	Other productivity shocks**	Prestige/ career/ age	Scientific fields	Regional share	Grp. size	Paper quality				
Peritz (1990)	✓	✓			✓				Descriptive analysis	No	No	No
Lewison and Dawson (1998)		✓				✓	✓	✓	Statistical analysis	No	No	No
Arora *et al.* (2000)	✓	✓		✓		✓	✓		OLS regression	No	No	No
Godin (2003)		✓			✓	✓		✓	Descriptive analysis	No	No	No
Payne and Siow (2003)		✓	✓					✓	Linear regression	No	No	No
Huffman and Evenson (2005)		✓				✓			Econometrics	Yes	No	No
Carayol and Matt (2006)		✓	✓	✓	✓				• Statistical analysis • OLS regression	No	No	No
Jacob and Lefgren (2007)		✓	✓		✓			✓	OLS regression	No	Yes	No
Crespi and Geuna (2008)		✓				✓		✓	Econometrics	No	Yes	No
Campbell *et al.* (2010)	✓	✓				✓		✓	Descriptive analysis	No	No	No
Beaudry and Clerk-Lamalice (2010)		✓				✓			Negative binomial regression	Yes	No	Yes
Beaudry and Allaoui (2012)		✓		✓		✓			Negative binomial regression	Yes	No	Yes

*No. of articles (or similar variables such authors/paper) have been considered as the dependent variable.
**E.g. The quality of institution, external shocks to schools, past publication pattern etc.

Apart from Crespi and Geuna (2008) who studied the time lag structure of the scientific output in 14 OECD countries, Jacob and Lefgren (2007) considered a couple of funding independent variables reflecting different periods of funding (including pre-funding period). Considering such a factor can help to better analyse the impact of funding on scientific output by distinguishing the impact of each period. However, Jacob and Lefgren (2007) used OLS regression on a limited set of variables while Crespi and Geuna (2008) used non-linear econometrics approach indicating the combined impact of the independent variables.

Another important point is that increasing the number of independent variables may augment the risk of having correlations among the variables. In order to calculate the net impact of funding on the scientific output more independent variables of different types are required. To overcome the risk of having correlations among the included variables one way is to increase the size of the population. Apart from the limited number of variables in the mentioned studies, sufficiently big population size has been also neglected by most of the researchers.

5. Research Gaps and Conclusion

Measuring and understanding the effect of funding is very critical. Financial investment may not necessarily be effective and may not result in higher scientific productivity. In order to have an efficient funding allocation, we need to understand the determinants of productive investment. If we know the marginal impact of research funding across disciplines, universities, researchers etc. it would be easier to plan a more efficient system.

Based on the literature reviewed, the relation among funding, collaboration and scientific production is still vague. Especially, our knowledge about the effects of funding on collaboration is very limited. Most of the studies considered a very limited scope such as the collaboration among the university professors or the cooperation between universities and industry. Moreover, no testing group has been considered in most of the studies. The other important issue is that we rarely found comprehensive analysis in the papers and whenever we see, for example, statistical analysis it is in the form of simplified linear regressions including limited number of independent variables and a narrow dataset. It is thus suggested to consider other forms of regressions such as non-linear equations or to add cross-relations between the independent variables which would allow analysing the combined effects of the independent variables while benefiting from the availability of the datasets of considerable sizes.

As mentioned in the previous sections, large amounts of money are being invested in the scientific development. To have a more accurate knowledge about the impact of funding on the scientific activities, various sources of funding, or various types of grants or funding programs should be evaluated separately. This issue has been rarely considered in the literature even though it can have important practical implications. Understanding the effectiveness of various granting strategies, programs and agencies will help the decision makers to distribute money more effectively.

In analysing the scientific collaboration, a weighted measure of publication is required. One simple way that is being used in the literature (e.g. Lee and Bozeman, 2005) is to divide a paper by the number of its co-authors and assign the resulted number to each of the co-authors as their production share. Other factors can be used when considering the weights, such as the average impact factor for the journals in each scientific field, average number of co-authors per paper, average citation etc. Therefore, a more comprehensive and accurate weighted measure can provide more valuable results. This issue can be generalised to other similar indicators (e.g. collaboration or team size indicators) since different scientific fields follow different patterns in publishing a paper or for collaboration. For example, in humanities most of the papers are single-authored while in engineering most of the papers have more than one author. Therefore, all the indicators should be normalised for various scientific fields and organisations.

In addition, network structure variables are important factors in evaluating the scientific collaboration or analysing the collaboration patterns. Considering such variables helps to study the collaboration more accurately. Although this issue was recently considered by Beaudry and Allaoui (2012), they have not focused on the effect of network structure variables and did not come with a comprehensive picture of the role of the network architecture. Moreover, in order to better analyse the impact of funding it is worth to consider the network variables in combination with other determinant factors (like past productivity of the researchers) to capture the impact in more detail.

To our knowledge, the role of funding and collaboration in scientific productivity has not been examined for the prominent individuals like star scientists and gatekeepers. Detecting the researchers who are playing an important role in scientific production and collaboration, and analysing their collaboration trends and funding patterns seem necessary in order to better understand their role in stimulating scientific activities and enhancing

research productivity in their communities. Such analysis will enable us to discover the characteristics of important players in scientific collaboration networks and will help policy makers to align their strategies in a way that improves the overall effectiveness of the funding programs and collaboration networks.

Nature of the science is becoming more complex and inter-disciplinary. Expertise from more and more disciplines is becoming necessary in order to produce knowledge. This should be reflected in funding policies of the governments and granting agencies, but no research so far has developed a clear link between the funding, multidisciplinary collaboration and knowledge production. Furthermore, in order to examine this issue the indicators for science complexity are needed. Such indicators will clearly denote a level of multi-disciplinarity and complexity for various research projects, and they will allow tracing their trends. Considering the limited sources of funding and the global research priorities of the country, policy makers can benefit from these indicators to design the most effective strategies.

More research has been done in the area of analysing funding impact on scientific output. Although few studies found positive impact (small in some cases, as discussed) of funding on scientific productivity, they mentioned that this positive effect would be due to the selection bias that may cause an overestimate of the impact of funding (Averch, 1987, 1988; Arora *et al.*, 2000; Godin, 2002). As an example, Arora *et al.* (2000) found that the researchers' characteristics such as their past productivity are related to their NSF grant. However, they declared that their estimates can be biased upward. Moreover, all the studies that evaluate the funding impact on scientific productivity at macro-level (government investment) are suffering from serious selection biases (Klette *et al.*, 2000).

Most of the studies have used bibliometrics or statistical methods for performing the analysis. Although bibliometrics is a simple and easy to use method, it is not an integrated approach since it considers too many assumptions that make the model very simplified (Ruegg, 2007). This could be also true for the statistical analysis and econometrics since the model is very limited and simplified in comparison with the real problem (Salter and Martin, 2001). Therefore, it is suggested to employ a variety of modern techniques such as data and text mining, social network analysis and 3-D visualisations to complement and validate the findings.

Considering the above, the need for a comprehensive study that covers several scientific fields and aspects, and an integrated evaluation that includes various state-of-the-art techniques is necessary especially for Canada since in comparison with other developed countries there was less effort in analysing the impact of funding on the scientific activities of different scientific fields and organisations. This can definitely help Canadian policy makers to adjust their strategies in a way that will lead the country to a better scientific position worldwide.

References

Adams, J, G Black, J Clemmons and P Stephan (2005). Scientific teams and institutional collaboration evidence from U.S. universities, 1981–1999. *Research Policy*, 34, 259–285.

Albrecht, C (2009). A bibliometric analysis of research publications funded partially by the Cancer Association of South Africa (CANSA) during a 10-year period (1994–2003). *South African Family Practice*, 73(1), 73–76.

Arora, A, PA David and A Gambardella (2000). Reputation and competence in publicly funded science: Estimating the effects on research group productivity. In *The Economics and Econometrics of Innovation*, 141–176. US: Springer.

Beaver, DD and R Rosen (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.

Beaudry, C and M Clerk-Lamalice (2010). Grants, contracts and networks: What influences biotechnology scientific production? In *Danish Research Unit for Industrial Dynamics (DRUID) Conference*, London (June 2010), 16–18.

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

Bourke, P and B Martin (1992). *Evaluating University Research Performance — What Approach? What Unit of Analysis?* Canberra: ANU and Brighton, SPRU.

Boyack, K and K Borner (2002). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology*, 54, 447–461.

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

Calero, C, TN van Leeuven and JW Tijssen (2005). Research networks of pharmaceutical firms: Geographical patterns of research collaboration within and between firms. In *Proceedings of the 10th International Conference on Scientometrics and Informetrics (ISSI)*, Vienna, Austria (July 2005), 310–315.

Campbell, D, I Labrosse, G Cote and E Archambault (2009). *Bibliometric Assessment of Research Funded by*

Genome Canada 1996–2007. Canada: Science-Metrix. Available at: http://www.genomecanada.ca/medias/PDF/EN/bibliometrics_assessment-of-Research-Funding.pdf [accessed on 1 July 2013].

Campbell, D, M Picard-Aitken, G Côté, J Caruso, R Valentim, S Edmonds and É Archambault (2010). Bibliometrics as a performance measurement tool for research evaluation: The case of research funded by the National Cancer Institute of Canada. *American Journal of Evaluation*, 31(1), 66–83.

Carayol, N and M Matt (2006). Individual and collective determinants of academic scientists' productivity. *Information Economics and Policy*, 18, 55–72.

Crespi, GA and A Geuna (2008). An empirical study of scientific production: A cross country analysis, 1981–2002. *Research Policy*, 37, 565–579.

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

De Solla Price, DJ (1963). *Little Science, Big Science*. New York: Columbia University Press.

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

Freedman, D, R Pisani and R Purves (1978). *Statistics*. New York: W. W. Norton.

García de Fanelli, A and M Estébanez (2007). Sistema nacional de innovación argentino: Estructura, grado de desarrollo y temas pendientes. In *Nuevos Documentos Cedes*, 31, 1–38.

Gingras, Y (1996). Bibliometric Analysis of Funded Research (A Feasibility Study). Report to the Program Evaluation Committee of NSERC.

Glanzel, W (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115.

Godin, B (2003). The impact of research grants on the productivity and quality of scientific research. (No. 2003). INRS Working Paper.

Gulbrandsen, M and JC Smeby (2005). Industry funding and university professors' research performance. *Research Policy*, 34, 932–950.

He, Z, X Geng and C Campbell-Hunt (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand University. *Research Policy*, 38, 306–317.

Heffner, AG (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, 3, 5–12.

Hirsch, JE (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.

Huffman, WE and RE Evenson (2005). New econometric evidence on agricultural total factor productivity determinants: Impact of funding composition. Iowa State University, Department of Economics, Working Paper 3029.

Jacob, B and L Lefgren (2007). The impact of research grant funding on scientific productivity. NBER Working Paper Series, Working Paper 13519.

Jaffe, A (1989). Real effects of academic research. *American Economic Review*, 79(5), 957–970.

Katz, J and B Martin (1997). What is research collaboration? *Research Policy*, 26, 1–18.

King, J (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13, 261–276.

Klette, TJ, J Moen and Z Griliches (2000). Do subsidies to commercial R&D reduce market failures? Microeconomic evaluation studies. *Research Policy*, 29, 471–495.

Lee, S and B Bozeman (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35, 673–702.

Lewison, G and G Dawson (1998). The effect of funding on the outputs of biomedical research. *Scientometrics*, 41(1), 17–27.

Leydesdorff, L and C Wagner (2009). Macro-level indicators of the relation between research funding and research output. *Journal of Infometrics*, 3(4), 353–362.

Lundberg, J, G Tomson, I Lundkvist, J Skar and M Brommels (2006). Collaboration uncovered: Exploring the adequacy of measuring university–industry collaboration through co-authorship and funding. *Scientometrics*, 69(3), 575–589.

Mansfield, E (1995). Academic research underlying industrial innovations: Sources, characteristics, and financing. *Review of Economics and Statistics*, 77(1), 55–65.

Martin, BR (2003). The changing social contract for science and the evolution of the university. In *Science and Innovation, Rethinking the Rationales for Funding and Governance*, A Geuna, AJ Salter and WE Steinmueller (eds.), Cheltenham: Edward Elgar.

Martin, BR and A Geuna (2003). University research evaluation and funding: An international comparison. *Minerva*, 41, 277–304.

Moed, HF and TN van Leeuwen (1996). Impact factors can mislead. *Nature*, 381, 186.

Okubo, Y (1997). Bibliometric indicators and analysis of research system: Methods and examples. OECD Working Papers.

Payne, AA and A Siow (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy*, 3(1).

Peritz, BC (1990). The citation impact of funded and unfunded research in economics. *Scientometrics*, 19, 199–206.

Porac, J, J Wade, H Fischer, J Brown, A Kanfer and G Bowker (2004). Human capital heterogeneity collaborative relationship, and publication patterns in a multidisciplinary

scientific alliance: A comparative case study of two scientific teams. *Research Policy*, 33, 661–678.

Rosenberg, N (1976). *Perspectives in Technology*. Cambridge: Cambridge University Press.

Rosenzweig, JS, SK Van Deusen, O Okpara, PA Datillo, WM Briggs and RH Birkhahn (2008). Authorship, collaboration, and predictors of extramural funding in the emergency medicine literature. *The American Journal of Emergency Medicine*, 26(1), 5–9.

Ruegg, R (2007). Overview of Evaluation Methods for R&D Programs. A Directory of Evaluation Methods Relevant to Technology Development Programs, U.S. Department of Energy.

Salter, AJ and BR Martin (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy*, 30, 509–532. SPRU-Science and Technology Policy Research, University of Sussex.

Savanur, K and R Srikanth (2009). Modified collaborative coefficient: A new measure for quantifying the degree of research collaboration. *Scientometrics*, 84, 365–371.

Seglen, PO (1992). The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628–638.

Seglen, PO (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314, 498–502.

Shapira, P and J Wang (2010). Follow the money: What was the impact of the nanotechnology funding boom of the past ten years? *Nature*, 468, 627–628.

Sonnenwald, DH (2008). Scientific collaboration. *Annual Review of Information Science & Technology*, 41(1), 643–681.

Subramanyam, K (1983). Bibliometric studies of research collaboration — A review. *Journal of Information Science*, 6(1), 33–38.

Thune, T (2007). University-industry collaboration: The network embeddedness approach. *Science and Public Policy*, 34(3), 158–168.

Tijssen, RJ (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, RJ, TN van Leeuwen and JC Korevaar (1996). Scientific publication actitivy of industry in the Netherlands. *Research Evaluation*, 6, 105–119.

Ubfal, D and A Maffioli (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, 40, 1269–1279.

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

Zucker, LG, MR Darby and MB Brewer (1998). Intellectual human capital and the birth of U.S. biotechnology enterprises. *American Economic Review*, 88(1), 290–306.

Scientometrics
DOI 10.1007/s11192-013-1069-6

Effect of collaboration network structure on knowledge creation and technological performance: the case of biotechnology in Canada

Hamidreza Eslami · Ashkan Ebadi · Andrea Schiffauerova

Received: 8 June 2013

Abstract Many of the novel ideas that lead to scientific publications or yield technological advances are the result of collaborations among scientists or inventors. Although various aspects of collaboration networks have been examined, the impact of many network characteristics on knowledge creation and innovation production remains unclear due to the inconsistency of the conclusions from various research studies. One such network structure, called small world, has recently attracted much theoretical attention as it has been suggested that it can enhance the information transmission efficiency among the network actors. However, the existing empirical studies have failed to provide consistent results regarding the effect of small-world network properties on network performance in terms of its scientific and technological productivity. In this paper, using the data on 29 years of journal publications and patents in the field of biotechnology in Canada, the network of scientists' collaboration activities has been constructed based on their co-authorships in scientific articles. Various structural properties of this network have been measured and the relationships between the network structure and knowledge creation, and quantity and quality of technological performance have been examined. We found that the structure of the co-authorship network of Canadian biotechnology scientists has a significant effect on the knowledge and innovation production, but it produced no impact on the quality of patents generated by these scientists.

Keywords Network structure · Small world · Innovation · Collaboration · Biotechnology · Canada

H. Eslami · A. Ebadi · A. Schiffauerova (✉)
Concordia Institute for Information Systems Engineering (CIISE), Concordia University, EV7.647, 1515 Ste-Catherine Street West, Montreal, QC H3G 2W1, Canada
e-mail: andrea@ciise.concordia.ca

Published online: 14 July 2013

Introduction

The concept of collective invention which consists in sharing of knowledge and information among a broad group of agents has been introduced by Allen (1983). It has been suggested that such sharing results in a fast knowledge accumulation and high invention rates. A number of examples of collective invention have been documented in the literature (e.g. Lamoreaux and Sokoloff 1997; von Hippel 1987; Schrader 1991; Saxenian 1994). Collective invention is driven by exchange and circulation of knowledge and information within networks formed by groups of socially connected individuals (Dahl and Pedersen 2004). The concept of collective invention is thus convenient for describing the dynamics of knowledge diffusion through innovation networks. Such networks are important highways of information and knowledge travelling among various scientists and inventors who collaborate and exchange knowledge in order to produce innovations and scientific knowledge. It has been suggested (Cowan and Jonard 2003) that the structure of the network over which the transmission of information takes place influences the extent of the diffusion and thus the technological potential of the firms. Consequently, the network architecture may be vitally important to the performance of the industry.

Much evidence supports the fact that some properties and the structure of networks influence the spread of knowledge (e.g. Abrahamson and Rosenkopf 1997; Schilling and Phelps 2007). Moreover, many other researchers like Cowan and Jonard (2001), Abrahamson and Rosenkopf (1997), Choi et al. (2010), Granovetter (1973) have emphasized the outstanding effect of network topologies on the performance of the system and the diffusion of innovation. Although the effects of the structural network properties on the network performance have been already studied, the findings of various studies sometimes include contradictory results. For example, clustering (cliquishness) is a property of a local network structure which refers to the likelihood that two vertices that are connected to a specific third vertex are also connected to one another. Cliquish networks have a tendency towards dense local neighborhoods, in which individual researchers or inventors are better interconnected with each other. Such networks exhibit a high transmission capacity, since a great amount of knowledge could be diffused rapidly (Burt 2001). Moreover, a high degree of cliquishness in an innovation network supports friendship and trust-building, and hence facilitates collaboration between innovators. Uzzi and Spiro (2005), Schilling and Phelps (2007) argue that higher cliquishness enhances system performance and knowledge diffusion. However, Cowan and Jonard (2003) point out the existence of negative effects of cliquishness stemming from the loss due to repetition, as the knowledge exchanged in highly cliquish neighborhoods is often redundant. Moreover, empirical findings of Fleming et al. (2007), Gilsing et al. (2008), He and Fallah (2009) confirm the negative impact of the higher cliquishness in the network on technological productivity. The role of a high degree of cliquishness in the innovation production is still not obvious and it is one of the objectives of this study to shed some light on the effects of various network properties.

One of the network structures that has recently attracted much theoretical attention is called small-world. In small-world networks high clustering could coexist with short path lengths. This creates an efficient network structure which is widespread in biological, social and man-made systems (Watts and Strogatz 1998). Many studies have been performed in various contexts to analyze the small-world effect in social networks (e.g. Travers and Milgram 1969; Choi et al. 2010; Latora and Marchiori 2001; Cowan and Jonard 2004; Fleming et al. 2007; Newman 2001a, b, c, 2004) and it has been suggested that the small-world property can enhance the information transmission efficiency among the network actors. As it is assumed that the level of the influence of network properties is

different in distinct industries (Feldman and Audretsch 1999), more studies elucidating this impact for various industrial sectors are needed.

The effects of the small-world network structure have already been investigated in the context of technological performance, but the results are still inconsistent. It has been proposed that the small-world network structure has an immense effect on enhancing the knowledge and innovation production (Watts 1999; Hargadon 2003; Cowan and Jonard 2004; Baum et al. 2003; Schilling and Phelps 2007; Uzzi and Spiro 2005). Nevertheless, the results of Fleming et al. (2007) failed to show any significant positive influence of the small-world structure on the technological performance of the network. Despite the significance of the collaboration and its network architecture for the creation of knowledge and innovation the amount of practical and empirical research performed within this theme is still scarce (Fleming et al. 2007).

To date, most of the existing research work in this area focused on the investigation of the network structure effects either within the realm of the scientific academic networks (e.g. Rumsey-Wairepo 2006) or within the industrial innovation networks (e.g. Schilling and Phelps 2007; Fleming et al. 2007; Beaudry and Schiffauerova 2011; He and Fallah 2009). In this study we decided to go further and not to stay only within the examination of the researchers' scientific collaborative activities and their scientific results, but to follow the researchers' activities further to see whether their scientific efforts could also lead to possible applications in the industrial context. By bringing together the data evidencing both scientific and technological aspects of the researchers' pursuits it becomes possible to analyze the effect of the network structure of scientific collaboration on the researchers' subsequent technological productivity. This is a very unique aspect of this research, because to our knowledge, no research work has examined the network properties within scientific networks while taking into consideration the researchers' patenting activities as well.

Furthermore, so far the attention of researchers was attracted mainly by the inter-firm networks, within which the effects of various properties have been examined (e.g. Cowan and Jonard 2003). However, our knowledge about the structure of knowledge flows and collaboration networks of individuals is still much more limited. This calls for further investigation of the network structure based on the collaboration among individual researchers.

Finally, despite the importance of networks in biotechnology the effect of collaboration in biotechnology sector has not been considered so far. It has been shown (Powell 1990) that the network ties existing among the researchers, inventors, universities and companies are among the most important factors that move the biotechnology sector forward and that they have a significant effect on the knowledge productivity. Liebeskind et al. (1996) confirmed that the dominance of scientists' social network is one of the major attributes of biotechnology sector. The collaboration networks of these scientists are considered to be significant drivers of research progress. Given the importance of the networks for the development of biotechnology sector it is surprising that no previous research work attempted to study the effects of the structure of knowledge flows among the biotechnology scientists. Our paper sheds first light on the effects of collaboration in biotechnology networks.

The main purpose of this research is to address the discussed research gaps by providing an empirical analysis of the collaboration network of individuals who carry out a scientific research in biotechnology in Canada. Our basic motivating questions are: What is the effect of Canadian biotechnology network structure on the research productivity of the scientists? What is the effect of this structure on the level of innovation productivity of

these researchers? What is its influence on the quality of technological output? Does Canadian biotechnology network resemble the small-world network structure? And does the small-world structure facilitate knowledge creation and technological performance of the researchers? The remainder of the paper will bring answer to these questions. It is organized as follows: "Methodology" section describes the data and methodology used in this paper. The empirical results and interpretations are provided in "Results" section, "Conclusions" section concludes with the findings of this study and the last "Limitations and future work" section discusses the limitations of this research while suggesting a possible future research avenues.

Methodology

The study has been conducted in two phases. During the first phase, the collaboration network of scientists has been constructed and social network analysis has been performed. A number of network indicators have been measured and collected as the input for the second phase. In the second phase, the association of the measured network properties with research productivity, innovativeness as well as innovation quality is examined. This phase encompasses a quantitative method using statistical analysis based on the data obtained from the previous phase. Moreover, the measured network indicators are used to indicate the small-world characteristics of the biotechnology network of Canada.

Data

To answer our research questions we created two databases, the database of Canadian biotechnology articles based on SCOPUS and the database of Canadian biotechnology patents extracted from the USPTO. We used the affiliations of the authors and the home addresses of the inventors to identify Canadian-based researchers and inventors. Only the articles and patents with at least one Canadian-based co-author or co-inventor have been extracted and included in our databases. In SCOPUS, we identified the biotechnology scientists using the set of biotechnology-related keywords, while in the USPTO we selected them according to OECD definition which is based on a group of carefully selected International Patent Codes (IPC).[1] Our final data covers 100,652 biotechnology articles published in the period from 1966 to 2005 and written by a total of 94,484 scholars. Out of these scientists, 5,013 collaborated on innovative projects and registered a total of 4,893 patents from 1971 to 2006. The database of the Canadian biotechnology articles has been employed to build the networks of collaborating scientists (based on the co-authorship of the articles), whose structure was then analyzed with Pajek, a social network analysis software. The Canadian biotechnology patent database has provided important information on the inventive productivity of the inventors.

We used the USPTO instead of the Canadian Intellectual Property Office (CIPO) database, since most of the Canadian biotechnology inventors prefer to protect their intellectual properties only in the US, or both in the US and in Canada. The main reason is the much larger biotechnology market of the United States within a relatively short geographical distance (Schiffauerova and Beaudry 2012; Niosi 2005; Aharonson et al. 2004).

[1] The OECD definition of biotechnology patents covers the following IPC classes: A01H1/00, A01H4/00, A61K38/00, A61K39/00, A61K48/00, C02F3/34, C07G(11/00, 13/00, 15/00), C07K(4/00, 14/00, 16/00, 17/00, 19/00), C12M, C12N, C12P, C12Q, C12S, G01N27/327, G01N33/(53*, 54*, 55*, 57*, 68, 74, 76, 78, 88, 92).

Co-authorship network of scientists

The collaboration networks were constructed based on the co-authorship of scholarly articles by individual scientists. We suppose that if the scientists co-author an article they have to be engaged in the collaboration research for some time. During this collaborative period it is assumed that information exchange happens among the scholars to a great extent (He and Fallah 2009). We have assumed that the length of the life of each link in the created network is 5 years. This assumption has been widely adopted in the previous research works (e.g. Baum et al. 2003; Fleming et al. 2007; He and Fallah 2009). Therefore, the publications were extracted for each five-year moving window from 1973 to 2005. This resulted in a total of 29 undirected one-mode networks.[2]

After the networks were constructed, their structure was analyzed with Pajek. The structural network properties that needed to be assessed in order to answer our research questions were measured and recorded. In order to analyze the effect of the network structure on the network productivity, first we need to quantify these concepts. To measure the productivity of the network, we need to define the necessary performance measures which are presented in the next section.

Variables in the model

We have studied the impact of the structural properties of scientists' collaboration network through three different performance measures (dependent variables) that were *research productivity of scholars*, *technological performance of scholars*, and *quality of the technological production*.

Research productivity of scholars (ARTCi)

To measure researchers' productivity, the number of the articles published by the scholars has been taken into consideration. According to Toutkoushian et al. (2003), the number of publications is the most common measure of scientists' research productivity. It has been widely used in the literature as a measure of research productivity (Centra 1983; Bland and Ruffin 1992; Taylor et al. 1984; Kuzhabekova 2011; Rumsey-Wairepo 2006). The notation used for this variable in this paper is *ARTCi*, which reflects the total number of articles published by individuals in year *i*.

Technological performance of scholars (Pi)

The patenting activity of the individuals was considered as a proxy for the second dependent variable, technological performance. The variable *Pi* is the indicator of number of patents in year *i*; where *i* ranges from 1973 to 2006. The number of patents has been a widely used indicator representing technological performance in many other research studies (e.g. Fleming et al. 2007; Schilling and Phelps 2007; He and Fallah 2009; Jaffe et al. 2000; Ahuja 2000).

[2] The first period consists of the scientists who published in 1973–1977, and the last one includes the ones who published in 2001–2005. Therefore, we have a total of 29 networks (Table 1).

Quality of the technological production (PQi)

The quality of the technological production was measured by the number of patent claims of the patents.[3] This value is noted as *PQi* and denotes the number of patent claims in year *i*.

We examined the effect of each structural network variables on the scientific and technological performance in the first year which follows the interval in which the network architecture was assessed. The reason for this is an assumption that the fruits of the scientists' 5-year collaborative period will be gathered only after this period has finished (Baum et al. 2003; Fleming et al. 2007; He and Fallah 2009). It usually takes time to publish an article or to register a patent. As a consequence, having a dependent variable calculated in year *i*, related independent variables are calculated for the networks constructed on the five-year snapshot from year $i - 5$ to $i - 1$. The independent variables are listed and described below.

Degree centralization (DC)

This variable is an indicator of variation in degree centrality of scientists in the network. The degree of a node is the number of links that are directly connected to the node (Wasserman and Faust 1994). The concept of centrality refers to the importance of the network actors in the process of information exchange in the network. When an actor is widely involved in the communications with other individuals, we can conclude that this actor plays an important role in the knowledge diffusion in the network. According to Wasserman and Faust (1994), this kind of involvement is called the centrality of the vertex. The centrality of a node can be analyzed from different aspects. One of the main indices for network analysis which is widely accepted in the literature (Schilling and Phelps 2007) is degree centrality. Degree centrality of a node measures the number of nodes that are directly connected to this node.

Clearly, the more a node is connected to other nodes, the more active it will be in the sense of information transfer, and consequently it will be more central. This value can be reflected in connectivity of the network that indicates the average degree of the whole network. According to He and Fallah (2009), when the degree centrality of network nodes varies more, the network will be more centralized. Therefore, we evaluate the centrality of the network by its degree centralization which is calculated as follows:

[3] Patent claims are a series of numbered expressions describing the invention in technical terms and defining the extent of the protection conferred by a patent (the legal scope of the patent). A high number of patent claims is an indication that an innovation is broader and has a greater potential profitability. It has been frequently suggested and empirically demonstrated (see for example Tong and Frame 1994) that the number of claims is significantly and consistently indicative of higher value patents. The conclusions of most of the papers on patent value reviewed by van Zeebroeck and van Pottelsberghe de la Potterie (2006) are supportive of positive association of the number of claims with patent value. Lanjouw and Schankerman (2004) have suggested that specifically in the biotechnology field the number of claims is the most important indicator of patent quality. Apart from patent claims, patent citations have been also considered as another quality index of patents (e.g. Jang et al. 2011; Fontana et al. 2009). However, one of the major limitations of patent citation data is that more citations could be added by the examiner without even informing the inventor. Alcácer and Gittelman (2006) found a very high magnitude for the examiners' citation effect where two-thirds of citations on an average patent are being inserted by examiners. This is being widely seen in the USPTO (Lukatch and Plasmans 2002). Hence, we used patent claims as the quality proxy of the patent in this paper.

$$DC = \frac{\text{Variation in degree centrality of nodes}}{\text{Maximum variation in degree centrality of a network of the same size}}$$

In general, the centrality variation measure is calculated by (Butts 2006):

$$C_*(G) = \sum_{v \in V} (\max_{v' \in V} C_*(v') - C_*(v))$$

where $C_*(G)$ denotes the graph centrality variation, $C_*(v')$shows the maximum centrality in the graph and $C_*(v)$ represents the centrality of each node. Clearly, the degree centrality measure should be normalized enabling the comparison of the graphs of different sizes. Therefore this value is divided by the highest possible variation in the degree centrality of a graph of the same size, and named degree centralization (de Nooy et al. 2005).

Betweenness centralization (BC)

Betweenness centrality takes into consideration the role of intermediary individuals, i.e. the scientists that lie on the paths connecting two nodes (Wasserman and Faust 1994). In other words, this measure evaluates the significance of a researcher as a connector between two other researchers as the existence of this connecting researcher can enhance the knowledge exchange between them. Therefore, the betweenness centrality of a node is defined as the proportion of all shortest paths between pairs of other nodes that include this node (de Nooy et al. 2005). This indicator shows the control of a node over the relations between the other individuals within the network and its impact on the information flow among them.

The variation in the betweenness centrality of nodes in a network is measured by betweenness centralization. According to de Nooy et al. (2005), this variable is calculated as below:

$$BC = \frac{\text{Variation in betweenness centrality of nodes}}{\text{Max variation in betweenness centrality of nodes in a network of same size}}$$

Clustering coefficient (CC)

This index evaluates the level of tendency of the nodes to cluster together. Local clustering coefficient of each node within a network is defined as:

$$CC_i = \frac{2z_i}{n_i(n_i - 1)}$$

where n_i is the number of direct neighbor nodes of node i and therefore the term $\frac{n_i(n_i-1)}{2}$ denotes the total number of possible links between node i's neighbors. z_i represents the total number of existing links between the n direct neighbors of the node i (Clements 2008). The average of the local clustering coefficients of all the nodes denotes the overall clustering coefficient of the entire network:

$$CC = \frac{1}{n} \sum_i CC_i$$

One of the main reasons for measuring the clustering of a network is to evaluate the small-world properties in the networks. However, the second small-world component, the path length, is not defined in the disconnected network, i.e. in the network where all the nodes are not directly or indirectly connected to each other and where isolated nodes may

exist. Therefore we measured both small world components, path length and clustering, only in the largest connected component of each network. Component of a network is a sub-network in which there is no isolated node and all the nodes are directly or indirectly connected to each other. In fact, the largest component is the largest fraction of the network where information exchange takes place.

Small-world (SW)

This measure is calculated by dividing the clustering coefficient by the average shortest path length of the network. Again, as was explained above, there is a limitation in calculating the small-world measure, since the small-world ratio cannot be defined in a disconnected graph. In this study, we follow the method used by most of the researchers in this area (Fleming et al. 2007; Uzzi and Spiro 2005; Kogut and Walker 2001; Baum et al. 2003) who consider only the largest connected component of the network for their analyses. Apart from the regression model, SW indicator was used for analyzing the small-world characteristics of the biotechnology network of Canada (Table 1).

Network size (Ln_Scts)

As new scientists join the network of biotechnology in Canada, there will be more chance for collaborations and as a result, more potential opportunity of knowledge exchange. This will clearly have an impact on the overall scientific and technological productivity. In order to account for the number of researchers in each network which can greatly affect our dependent variables, we used a control variable for the size of the network that accounts for the growth of the network. The variable *Ln_Scts* takes natural log values in order to compress the data scale.

Results

The first aspect of the Canadian biotechnology network being observed is its size, i.e. the total number of scientists that are engaged in at least one research activity in the corresponding period of time (Fig. 1). The sudden jump in the population growth of scientists in 1980s has been explained as the result of popularity of internet in different research areas (Munn-Venn and Mitchell 2005).

Although the degree centralization drops in the first years (Fig. 2), the value of degree centralization remains below 0.01 during the whole studied period. This means that the distribution of the links among scientists is almost the same for the whole network, resulting in a homogenous growth of collaborative opportunity for all scientists in the network. As it is depicted in Fig. 3, the clustering coefficient first increases and then remains relatively constant in final periods.

Figure 4 shows the actual path length of the Canadian biotechnology networks. It is observed that the path length is almost constant during 1977 to 1983. It can be said that in this period the network size is relatively large; hence the nodes have become more separated having relatively a large path length (Albert and Barabási 2002), meaning that more nodes have to be traversed to reach other scientists. After 1983, the shortest path starts to decrease. The declining trend of this measure is considered to be one of the main indicators of small-world phenomenon in large networks. The shortest path reduction proves that the number of ties between scientists has increased and there are more links created both

Table 1 Small-world characteristics for the Canadian biotechnology networks

Time intervals	Number of nodes	Path length		Clustering coefficient		SW
		PL_a	PL_r	CC_a	CC_r	
1973–1977	1169	16.11	5.61	0.741	0.0030	85.66
1974–1978	1970	14.59	6.01	0.743	0.0018	170.86
1975–1979	2407	15.68	6.12	0.753	0.0015	198.47
1976–1980	2645	14.20	6.13	0.763	0.0014	240.92
1977–1981	2808	15.38	6.19	0.758	0.0013	237.94
1978–1982	3089	16.09	6.26	0.757	0.0012	252.19
1979–1983	3541	17.13	6.35	0.764	0.0010	277.34
1980–1984	3891	15.10	6.42	0.758	0.0009	346.63
1981–1985	4831	14.25	6.57	0.762	0.0008	466.88
1982–1986	5206	12.82	6.56	0.759	0.0007	547.92
1983–1987	5915	11.72	6.21	0.769	0.0007	595.59
1984–1988	7922	10.02	5.61	0.789	0.0006	704.93
1985–1989	9909	8.75	5.35	0.796	0.0006	862.24
1986–1990	11760	8.04	5.17	0.800	0.0005	988.64
1987–1991	14131	7.63	5.08	0.802	0.0005	1150.79
1988–1992	16804	7.17	5.02	0.803	0.0004	1358.08
1989–1993	18980	6.97	4.99	0.802	0.0004	1514.06
1990–1994	21297	6.73	5.00	0.801	0.0003	1726.10
1991–1995	23229	6.57	4.97	0.801	0.0003	1860.51
1992–1996	25954	6.43	4.98	0.794	0.0003	2077.63
1993–1997	27980	6.29	5.00	0.789	0.0003	2260.65
1994–1998	29266	6.14	4.88	0.785	0.0003	2222.68
1995–1999	30250	6.03	4.86	0.779	0.0003	2275.84
1996–2000	30772	5.93	4.73	0.777	0.0003	2146.59
1997–2001	32087	5.81	4.68	0.782	0.0003	2202.37
1998–2002	33466	5.71	4.55	0.792	0.0003	2130.35
1999–2003	34541	5.67	4.56	0.791	0.0003	2226.03
2000–2004	35640	5.59	4.52	0.792	0.0003	2248.34
2001–2005	37730	5.47	4.54	0.790	0.0003	2422.75

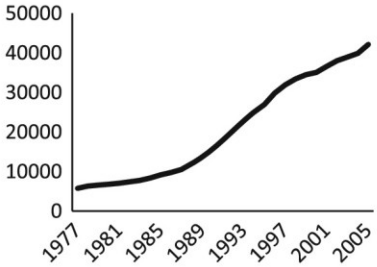


Fig. 1 Historical trend of network size from 1977 to 2005

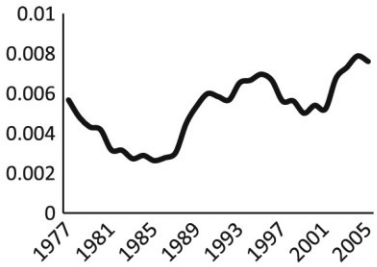


Fig. 2 Historical trend of degree centralization from 1977 to 2005

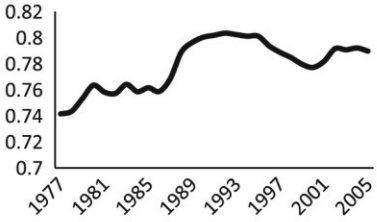


Fig. 3 Historical trend of clustering coefficient from 1977 to 2005

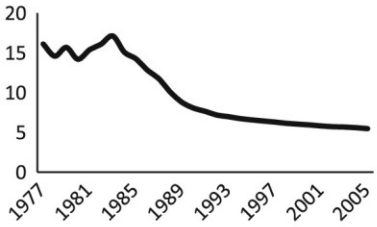


Fig. 4 Historical trend of actual values of path length from 1977 to 2005

among the scientists already existing in the networks and between the new individuals entering the networks and the existing ones (Clements 2008).

The results of path length measurements reveal an important outcome: interestingly, the values of path length converge to 6 in the later periods. This implies that the average distance between individuals is around 6, which is consistent with the results of Milgram's (1967) who first introduced the notion of the small-world structure. Based on his empirical study, to reach a person who is unknown to an individual, on average only six intermediates are needed.

Small world characteristic

According to Albert and Barabási (2002), the small-world networks are often large in size, but despite their size they still exhibit fairly short path lengths and high cliquishness. In order to measure to what extent the structure of our network resembles the structure of a small-world network, we follow the approach of Watts (1999) in which the path length and clustering coefficient are modified first to be used in the small-world equation and then use them to gain the value of small-world characteristic for each network based on the method employed in several previous studies (Davis et al. 2003; Kogut and Walker 2001; Baum et al. 2003).

The results are presented in Table 1. There is no critical index for the small-world measure. Besides, it is implied that when the size of the network increases, the critical value for the small-world should increase (Baum et al. 2003). Therefore, the common

Table 2 A comparison of previously studied small-world networks (Kogut and Walker 2001; Albert and Barabási 2002)

Network	CC_a/CC_r	PL_a/PL_r	SW	Network size	Ref.
Ythan estuary food web	3.67	1.08	3.4	134	Montoya and Sole (2002)
E. coli substrate graph	12.31	0.96	12.83	282	Fell and Wagner (2000)
E. coli, reaction graph	6.55	1.32	4.96	315	Fell and Wagner (2000)
Power grid	16	1.51	10.6	4941	Watts and Strogatz (1998)
NCSTRL co-authorship	1653.34	1.16	1425.3	11994	Newman (2001c)
Words, synonyms	1166.7	1.18	988.73	22311	Yook et al. (2001)
LANL co-authorship	2388.9	1.23	1942.2	52909	Newman (2001c)
SPIRES co-authorship	242	1.89	128.05	56627	Newman (2001c)
Math co-authorship	10925.93	1.16	9418.91	70975	Barabási et al. (2002)
MEDLINE co-authorship	6000	0.94	6382.98	1520251	Newman (2001c)

procedure (Albert and Barabási 2002; Davis et al. 2003; Kogut and Walker 2001; Baum et al. 2003) to find out whether the networks exhibit small-world properties or not consists in comparing their small-world values to the networks previously studied in the literature. The list of corresponding values of some of the previously identified small-world networks are summarized in Table 2.

By comparing the SW values measured on the networks of this study with the values measured on the networks of similar sizes in the prior research, it can be concluded that the Canadian biotechnology network evidently represents the small-world properties with respect to its high SW values.

For example, we can observe a small-world measure of 12.83 for the network of *E. coli* substrate graph studied by Fell and Wagner (2000), whereas the network of our study with similar size shows the value of 30.69 for this variable; or the network of SPIRES co-authorship analyzed by Newman (2001c) demonstrates a SW value of 1,942.2, whereas the network of similar size in our study has the value of 2,422.75 for this variable. These comparisons and the increasing trend of the SW values for our networks leave no doubt that they have small-world structure.

Correlation analysis

In order to be able to determine the association between dependent and independent variables, a correlation analysis for the significance level of 99 % was performed by STATA 11 between the independent variables of the regression model. The results of the correlation analysis are shown in Table 3.

Table 3 Correlation analysis of the Canadian biotechnology network

	DC	BC	CC	SW
DC	1.0000			
BC	−0.3350	1.0000		
CC	0.4483	−0.2857	1.0000	
SW	0.4729	−0.1429	0.4532	1.0000

As the results show, the correlation measures between the independent variables of the model (i.e. degree centralization, betweenness centralization, clustering coefficient, and small-world variable) are not significant.

Regression analysis

Three separate regression models are developed (one for each of the dependent variables). According to Hausman et al. (1984), in order to provide a natural baseline for a count measure, the regression model of choice is a Poisson model. Since our three dependent variables, i.e. the total number of articles, patents and patent claims (patent quality), are count measures, the best matching regression model would be Poisson. However, the primary assumption for Poisson model is that it accepts no heterogeneity in the data, which means that variance and mean of the sample should be the same (Coleman 1981). In reality, however, it is rare to satisfy the Poisson assumption on the actual distribution of a natural phenomenon, because most of the time an over-dispersion or under-dispersion is detected in the sample data. This causes the Poisson model to underestimate or overestimate the standard errors and thus results in misleading estimates for the statistical significance of variables (Coleman 1981). Hausman et al. (1984) suggest correcting the estimates by using negative binomial regression models instead of Poisson in order to obtain robust standard errors. To select the best regression model for each of the three dependent variables, we did the likelihood ratio test. Since the overdispersion coefficient (α) for total number of patents and total number of articles reported a small value close to zero we accepted the null hypothesis that Poisson is a better estimation model in their cases. However, the overdispersion coefficient (α) value for patent quality was relatively significant, which suggested that the negative binomial regression model is a better estimator in this case.

The observation of the regression coefficients for the impact of the Canadian biotechnology network structural properties on the network's research performance in terms of number of articles is presented in Table 4. Since the p values reported for all the independent variables (except for the betweenness centralization *BC*) are <0.01, they are considered as significant predictors in the knowledge productivity of the following year. Moreover, the model was also tested for robustness and consistency of the results.

Table 4 Poisson regression results for total number of articles model

Poisson regression					Number of obs = 29	
					LR $\chi^2(5) = 1551333.75$	
					Prob > $\chi^2 = 0.0000$	
Log likelihood = −271.57972					Pseudo $R^2 = 0.9965$	
ARTC	Coef.	SE	*z*	*P* > |z|	95 % CI	
DC	25.26977	2.280177	11.08	0.000	29.73883	20.8007
BC	0.1490192	0.2134787	0.70	0.485	−0.2693913	0.5674298
CC	−5.436369	0.2802158	−19.40	0.000	−5.985582	−4.887156
SW	0.0008818	0.0002505	3.52	0.000	0.0003907	0.0013728
Ln_Scts	0.994302	0.0364738	27.26	0.000	0.9228147	1.065789
_cons	3.918968	0.1974064	19.85	0.000	3.532059	4.305878

SE standard error, *CI* confidence interval

In addition, to verify the stability of the model we used bootstrapping re-sampling method to bootstrap the standard errors of the parameter estimates by performing 50 replications. The results of the mentioned tests show that apart from *BC*, *SW* becomes also insignificant. This does not affect the original findings since as it is shown in Table 4 the coefficient of *SW* is very small.

Table 4 reports a very strong negative influence of the degree centralization (*DC*) on the productivity. We can conclude that the central structure of the network reduces the overall knowledge spillovers among the scientists, resulting in less productivity in the upcoming year. This negative effect is expected since authors with high degree centrality can influence the knowledge diffusion by withholding the transmission of information (Bavelas 1950; Chung and Hossain 2009; Freeman 1979; Leavitt 1951). The betweenness centralization (*BC*) regressor in the model is not significant, which means that we cannot make any conclusion regarding its effect on the Canadian biotechnology scientists' network.

The results for the small-world indicator suggest a negative influence of clustering coefficient (*CC*) and a small positive impact of small-world (*SW*) on the research productivity of the subsequent year. For small-world, this result was expected since an increased average small-world measure of a network may result in faster and easier exchange of knowledge, which is in accordance with the major conclusions of other researchers (e.g. Latora and Marchiori 2001; Kogut and Walker 2001; Cowan and Jonard 2004). However, it should be noticed that the small-world effect is very small (0.0008818).

The result of the regression analysis for the model built on the technological productivity of individuals (total number of patents) is shown in Table 5. All of the independent variables in this model have a significant impact on the patenting of the subsequent year (except *Ln_Scts* which is our control variable).

Similarly as was observed in the articles model, Table 5 indicates a very strong negative influence of the degree centralization (*DC*) on the total number of patents. It can be said that the central structure of the network reduces the overall technological diffusion among the innovators, resulting in less patenting productivity in the coming year. This negative effect can be explained similarly as in the articles model, i.e. authors with high degree centrality can withhold the transmission of information and thus negatively influence the knowledge and technological diffusion.

Table 5 Poisson regression results for total number of patents model

Poisson regression					Number of obs = 29	
					LR $\chi^2(5) = 3638.68$	
					Prob > $\chi^2 = 0.0000$	
Log likelihood = −227.53601					Pseudo $R^2 = 0.8888$	
P	Coef.	SE	*z*	*P* > |z|	95 % CI	
DC	−73.5232	23.07165	−3.19	0.001	−118.7428	−28.30359
BC	−9.561112	2.647392	−3.61	0.000	−14.74991	−4.372319
CC	15.26929	2.89558	5.27	0.000	9.594057	20.94452
SW	0.0073319	0.0023938	3.06	0.002	0.0026402	0.0120235
Ln_Scts	0.453925	0.35317	1.29	0.199	−0.2382755	1.146125
_cons	−11.93914	2.272292	−5.25	0.000	−16.39275	−7.485525

SE standard error, *CI* confidence interval

Betweenness centralization shows a negative effect on patenting activity as well. The fact that a lot of information passes through the authors with high betweenness centralities suggests that these individuals have a power to control the flow of innovation through the network. Our results show that the presence of this power in the network does not support its technological performance. It is thus possible that the inventors who occupy key central positions in the network have a lower tendency to diffuse innovative ideas among others. Consequently we conclude that the more homogeneous the intermediary roles of the individuals are the less of such power exists in the network and the better innovation diffuses among the inventors. This subsequently results in higher patent productivity. These findings are partially in accordance with the results of Gilsing et al. (2008) who concluded that after a certain threshold higher betweenness centralities result in a negative impact on the technological productivity of the network.

Again, similarly as in the articles model, the positive small-world effect on network technological productivity is slightly supported in this model as well.

The patents model shows the positive influence of the clustering coefficient, which is an opposite result when compared to the articles model. The role of clustering in networks has been analyzed by several researchers, whose conclusions are not consistent, because both positive and negative effects have been reported. Our articles model supports the finding that the high clustering of the co-authorship network limits the knowledge creation, which is consistent with the outcomes of for example Fleming (2007), Gilsing et al. (2008), He and Fallah (2009). However, the patents model supports the studies claiming the positive effect of clustering on the productivity of network actors (e.g. Schilling and Phelps 2007). These are very interesting results showing that the way the scientists collaborate among themselves has a quite distinct impact on their scientific and technological performances. Communities of researchers working in clustered neighborhoods will be less productive in terms of the journal papers, but they will be more successful in producing potentially commercializable inventions. Further research is needed to better understand and explain the factors and circumstances under which clustering plays a positive role and the ones which make clustering an element impeding scientific and technological performance. This work has shed some light on the reasons behind the inconsistency of the results of various studies on clustering. We propose that the scientific and technological aspects of the collaboration should be considered and studied together.

Finally, Table 6 illustrates the results for our last model, i.e. the negative binomial regression model developed for evaluation of the effect of the network structure on the patent quality (measured by the total number of patent claims). It shows no significant impact of the network indicators, since all the calculated p values are very much larger than the critical value of 0.01. This fact strongly implies that the network of article co-authorships does not give us enough evidence to assess the impact of the network structure of scientists' relationship on the quality of registered patents in Canadian biotechnology sector.

Table 7 depicts the number of observations and the significance levels for the three aforementioned models (*ARTC*, *P*, *PQ*). The results confirm that at the significance level of 0.1 (highlighted by + in the table) just two independent variables (*DC* and *Ln_Scts*) of the *PQ* model have significant correlations. However, we ignore these correlations since no significant impact has been found for the *PQ* model (Table 6). The independent variables of the other two models (*ARTC* and *P*) have no significant correlation at the significance level of 0.1.

Table 6 Negative binomial regression results for the total number of patent claims model

Negative binomial regression					Number of obs = 29	
					LR $\chi^2(5) = 84.91$	
Dispersion = mean					Prob > $\chi^2 = 0.0000$	
Log likelihood = −211.31581					Pseudo $R^2 = 0.1673$	
PQ	Coef.	SE	*z*	*P* > |z|	95 % CI	
DC	−165.6366	82.4919	−2.01	0.045	−327.3178	−3.955498
BC	5.25473	5.6171	0.94	0.350	−5.754584	16.26404
CC	6.844759	8.757269	0.78	0.434	−10.31917	24.00869
SW	−0.0030488	0.007889	−0.39	0.699	−0.0185111	0.0124134
Ln_Scts	2.354932	1.12505	2.09	0.036	0.1498741	4.559989
_cons	−19.98218	5.422574	−3.68	0.000	−30.61023	−9.354129
/lnalpha	−2.59509	0.2668259			−3.118059	−2.07212
Alpha	0.0746392	0.0199157			0.044243	0.1259185

SE standard error, *CI* confidence interval

Table 7 Significance levels

	(1) ARTC	(2) P	(3) PQ
DC	−25.27* (−11.08)	−73.52* (−3.19)	−165.6+ (−2.01)
BC	0.149 (0.70)	−9.561* (−3.61)	5.255 (0.94)
CC	−5.436* (−19.40)	15.27* (5.27)	6.845 (0.78)
SM	0.000882* (3.52)	0.00733* (3.06)	−0.00305 (−0.39)
Ln_Scts	0 0.994* (27.26)	0.454 (1.29)	2.355+ (2.09)
_cons	3 0.919* (19.85)	−11.94* (−5.25)	−19.98* (−3.68)
lnalpha_cons			−2.595* (−9.73)
#obs.	29	29	29

t statistics in parentheses

+ $p < 0.1$; * $p < 0.01$

Conclusions

This study explores the network of biotechnology scientists and inventors in Canada and examines the relationship between various structural properties of the network (particularly small-world properties) and the research and technological performance of scientists and inventors within the network. According to the results, the analyzed network showed significant small-world properties in any aspect i.e. the path length of the networks are very close to the ones of random networks; clustering coefficients of the networks are much

larger than the corresponding values of clustering coefficients approximated for the random networks; and finally, the small-world ratios are great or even larger than the corresponding values of previously studied networks.

Another interesting finding of this study deals with the Milgram's (1967) six degrees of separation. Our results strongly support Milgram's finding since the separation degree between scientists converge to six in the networks built on the later time periods. Therefore, as our networks demonstrate more small-world characteristics, the number of intermediaries between individuals become closer to six.

Concerning the relationship between the structural properties of the co-authorship network of Canadian biotechnology scientists and their knowledge output, our results proved that there is a significant association between the way the scientists are interconnected among themselves when collaborating on their research papers and the number of publications and patents arising from these collaborations. However, based on the results, there is no great association between the pattern of knowledge exchange among the collaborating scientists in the network and the quality of network's innovation productivity, assessed by the number of patent claims.

The articles and patents models' outcomes were in accordance with the findings of some previous studies in which the effect of network structure on its productivity was assessed (Latora and Marchiori 2001; Cowan and Jonard 2004; Fleming et al. 2007; McFadyen et al. 2009). Hence, the common hypothesis which states small-world properties enhance the knowledge creation is supported. The results show positive impact of small-world structure on the knowledge and innovation productivity of Canadian biotechnology sector.

This work is to our knowledge the first which has examined the network properties within scientific networks and studied their impact on the researchers' both scientific and technological productivity. It has proved to be a fruitful research approach. By bringing together the data evidencing both scientific and technological aspects of the researchers' pursuits we were able to distinguish between the features of collaborative activities which are beneficial for the researchers' scientific productivity from those which lead to their enhanced technological performance.

One of the interesting findings which this approach enabled us to obtain is the distinct effect of clustering observed for the scientific and technological production of the researchers. As mentioned previously, highly cliquish network structure does not enhance the knowledge creation. One of the reasons suggested by Cowan and Jonard (2003) is that such structure of scientific collaboration involves a loss due to repetition. This means that since individual researchers in the cliquish neighborhoods collaborate within closely related communities they are better interconnected with each other and the knowledge they exchange is often redundant. This leads to the subsequent decline in the overall knowledge productivity of scientists. Moreover, in highly cliquish networks the distances between the researchers are usually long, and long path length has a negative effect on the scientific knowledge productivity, because such network architecture decreases the efficiency of knowledge transfer. The long path lengths usually co-exist with most of the cliquish networks, but the exception is here the small world network structure where clustered neighborhoods in fact appear together with short path distances among the network nodes. Our results thus show support both for the negative influence of the clustered collaborative network structure on the scientific productivity in the network, and for the fact that there exists an exception to this which emerges when the positive path length effect overweighs the negative effect of high cliquishness—in the small world network structures.

On the other hand, we find that there are beneficial properties of the clustered network structure. The researchers are well interconnected and closely related within the highly

clustered communities, and as a consequence, a high degree of cliquishness in such networks supports friendship, confidence and trust-building. These are important factors when it comes to the commercial pursuits of the researchers' discoveries, because the co-partnership in potentially profit making affairs requires a higher level of confidence and trust, as opposed to the co-publishing of scientific articles. This may explain the different effect of clustering obtained for the scientific and technological production of the researchers. The high level of cliquishness in scientific communities hence hinders the knowledge productivity because such network structure does not provide most efficient conditions for transfer and creation of knowledge, but it facilitates the efforts of scientists leading to possible applications in the industrial context due to the increased trust and confidence existing in such research communities.

Moreover, our approach allowed us to observe the distinct effects of betweenness centralization on the knowledge and innovation creation. In fact, if we had measured only the impact of the scientific collaboration structure on the scientific performance of the researchers we would not have encountered any significant effect of betweenness in our model. It however becomes significantly negative in the patents model where its effect on the technological productivity is estimated. This suggests that the fact that there are individuals with a great power capable of controlling the flow of knowledge through the network does not constitute any real hindrance for the knowledge transmission and creation. Nevertheless, when the research efforts may potentially end up as profitable industrial applications, the potential power game changes the atmosphere in the network. In this case, a more homogeneous power distribution throughout the network provides a more productive environment for the researchers working on the patentable ideas and inventions.

Limitations and future work

We were exposed to some limitations in our analysis. First of them is related to the sample size. For years there has been a discussion on the accurate and sufficient sample size for the regression analysis, and various rules of thumb and methods for its determination have been proposed. Historically, the most common rule of thumb indicates that at least 10 samples are required for each independent variable (Maxwell 2000; Harris 1985), but other ratios were proposed as well. For example, according to Wampold and Freund (1987), the adequate sample size can be smaller than 10:1 ratio, and specifically for the multiple regressions, Hair et al. (1998) proposed 5:1 ratio as a general rule, which would be satisfied in our work. One of the main reasons for the discrepancies in the rule of thumb recommendations is that the applications of regressions are numerous. Choosing the appropriate sample size is thus a more complex issue. Even for the same statistical analyses some researchers may agree on different adequate sample sizes if they take into consideration various other criteria for the models (Knofczynski and Mundfrom 2008). Using a Monte Carlo simulation technique, Knofczynski and Mundfrom (2008) proposed the adequate sample size for various scenarios. We used their technique and while considering our model characteristics and number of variables, we found that our sample size would be considered sufficient. Moreover, in order to account for the possible sample size limitation we carried out additional validation techniques to test the model stability via bootstrapping. Considering the validation results, our model characteristics and the nature of the network analysis studies, we are convinced that that our model has a good predictive performance.

Second limitation involves our ability to calculate the small-world indicators only on the largest connected component of the network, and not on the full network containing all the nodes. Although there have been some recommendations as to how to resolve this issue

(e.g. Schilling and Phelps 2007), the solutions are usually applicable when the special purpose software for social network analysis is used. Therefore, future work can address this issue by coding a special purpose program to calculate the small-world ratio over the whole network.

Moreover, we were not able to assess the relationship between the structural properties of the articles co-authorship network and the quality of publications produced by the network members. The most commonly accepted measure for the quality of the research articles is the number of citations which each individual article receives from other citing research papers. However, since our data did not include this information, we could not examine this relationship. Apart from this, another interesting factor to be considered is the strength of the association between scientists. Multiple connections between individuals were considered as a single link in our networks. It is thus proposed here to consider the number of articles coauthored by the scientists during each time interval as a measure of the strength of their relationship. Hence, further research is needed to investigate this issue.

Many important factors that affect our dependent variables could not be incorporated in the models. For example, even though our methodology is able to detect and analyze the research collaborations leading to some tangible output (article, patent), the informal relationships that exist among scientists were completely neglected. These types of connections are never recorded and thus cannot be quantified, but there are certainly some knowledge exchanges occurring during such associations that could affect the network performance.

Finally, we could not use patent citations, because we do not have the patent citation data in our database. Patent citations have been used as an indicator of the quality of patents (e.g. Jaffe et al. 1993; Hu et al. 2003), but as we suggested previously the patent examiners' effect is being widely seen in the USPTO database (Lukatch and Plasmans 2002). We have considered the number of patent claims as the patent quality proxy instead. However, a future work can analyze the models using both patents citations and patent claims as patent quality measures and compare the results.

References

Abrahamson, E., & Rosenkopf, L. (1997). Social network effects on the extent of innovation diffusion: A computer simulation. *Organization Science, 8*, 289–309.

Aharonson, B., Baum, J., & Feldman, M. (2004). Industrial clustering and the returns to inventive activity: Canadian biotechnology firms, 1991–2000.

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly, 45*, 425–455.

Albert, R., & Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47.

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics, 88*(4), 774–779.

Allen, T. (1983). Collective invention. *Journal of Economic Behavior and Organization, 4*, 1–24.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3–4), 590–614.

Baum, J., Shipilov, A., & Rowley, T. (2003). Where do small worlds come from? *Industrial and Corporate Change, 12*(4), 697.

Bavelas, A. (1950). Communication patterns in task oriented groups. *Journal of the Acoustical Society of America, 22*, 271–282.

Beaudry, C., & Schiffauerova, A. (2011). Impacts of collaboration and network indicators on patent quality: The case of Canadian nanotechnology innovation. *European Management Journal, 29*(5), 362–376.

Bland, C. J., & Ruffin, M. T. (1992). Characteristics of a productive research environment: Literature review. *Academic Medicine, 67*(6), 385–397.
Burt, R. (2001). Structural holes versus network closure as social capital. *Social capital: theory and research*. Amsterdam: Kluwer Academic Publishers.
Butts, C. T. (2006). Exact bounds for degree centralization. *Social Networks, 28*(4), 283–296.
Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education, 18*(4), 379–389.
Choi, H., Kim, S. H., & Lee, J. (2010). Role of network structure and network effects in diffusion of innovations. *Industrial Marketing Management, 39*(1), 170–177.
Chung, K., & Hossain, L. (2009). Measuring performance of knowledge-intensive workgroups through social networks. *Project Management Journal, 40*(2), 34–58.
Clements, M. (2008). Patenting at universities in the United States: A network analysis of the complexities of domestic and international university patenting activities.
Coleman, J. (1981). *Longitudinal data analysis*. New York: Basic Books. 102-081-920.
Cowan, R., & Jonard, N. (2001). *Knowledge creation, knowledge diffusion and network structure* (pp. 327–343). Berlin: Springer.
Cowan, R., & Jonard, N. (2003). The dynamics of collective invention. *Journal of Economic Behavior & Organization, 52*(4), 513–532.
Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control, 28*(8), 1557–1575.
Dahl, M., & Pedersen, C. (2004). Knowledge flows through informal contacts in industrial clusters: myth or reality? *Research Policy, 33*, 1673–1686.
Davis, G., Yoo, M., & Baker, W. (2003). The small world of the American corporate elite, 1982–2001. *Strategic Organ, 1*(3), 301–326.
de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.
Feldman, M., & Audretsch, D. (1999). Innovation in cities: Science-based diversity, specialization and localized competition. *European Economic Review, 43*, 409–429.
Fell, D., & Wagner, A. (2000). The small world of metabolism. *Nature Biotechnology, 18*, 1121–1122.
Fleming, L., King, C., & Juda, A. (2007). Small worlds and regional innovation. *Organization Science, 18*(6), 938–954.
Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. *Network Strategy, 18*(4), 311–336.
Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.
Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy, 37*, 1717–1731.
Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*, 1360–1380.
Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). New Jersey: Prentice Hall.
Hargadon, A. (2003). *How breakthroughs happen: the surprising truth about how companies innovate*. Boston: Harvard Business School Press.
Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic Press.
Hausman, J., Hall, B., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica, 52*, 909–937.
He, J., & Fallah, M. H. (2009). Is inventor network structure a predictor of cluster evolution? *Technological Forecasting and Social Change, 76*(1), 91–106.
Hu, A., Adam, G. Z., & Jaffe, B. (2003). Patent citations and international knowledge flow: The cases of Korea and Taiwan. *International Journal of Industrial Organization, 21*(6), 849–880.
Jaffe, A., Trajtenberg, M., & Fogarty, M. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review, 90*(2), 215–218.
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics, 108*(3), 577–598.
Jang, S. L., Yu, Y. C., & Wang, T. Y. (2011). Emerging firms in an emerging field: an analysis of patent citations in electronic-paper display technology. *Scientometrics, 89*(1), 259–272.
Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement, 68*(3), 431–442.
Kogut, B., & Walker, G. (2001). The small world of Germany and the durability of national networks. *American Sociological Review, 66*, 317–335.

Kuzhabekova, A. (2011). Impact of co-authorship strategies on research productivity: A social-network analysis of publications in Russian cardiology (Doctoral dissertation, University of Minnesota, 2011).
Lamoreaux, N., & Sokoloff, K. (1997). Location and technological change in the American glass industry during the late nineteenth and early twentieth century. *5938*.
Lanjouw, J., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *Economic Journal, 114*, 441–465.
Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*(19), 198701.
Leavitt, H. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology, 46*(1), 38–50.
Liebeskind, J., Oliver, A., Zucker, L., & Brewer, M. (1996). Social networks, learning, and flexibility: Sourcing scientific knowledge in new Biotechnology firms. *Organization Science, 7*(4), 428–443.
Lukatch, R., & Plasmans, J. (2002). Measuring knowledge spillovers using patent citations: Evidence from Belgian firms' data.
Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*(4), 434–458.
McFadyen, M. A., Semadeni, M., & Cannella, A. A. (2009). Value of strong ties to disconnected others: Examining knowledge creation in biomedicine. *Organization Science, 20*(3), 552–564.
Milgram, S. (1967). The small world problem. *Psychology Today, 2*(1), 60–67.
Munn-Venn, T., & Mitchell, P. (2005). *Biotechnology in Canada: A technology platform for growth*. Conference Board of Canada.
Montoya, J. M., & Sole, R. V. (2002). Small world patterns in food webs. *Journal of Theoretical Biology, 214*(3), 405–412.
Newman, M. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review, 64*(1), 016131.
Newman, M. (2001b). *The structure of scientific collaboration networks: Proceedings of the National Academy of Sciences, 98(2), 404.*
Newman, M. E. (2001c). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.
Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1), 5200.*
Niosi, J. (2005). *Canada's regional innovation systems. The science-based industries*. Montreal: McGill-Queen's University Press.
Powell, W. (1990). Neither market nor hierarchy: Network forms of organization. *Research in Organizational Behavior, 12*, 295–336.
Rumsey-Wairepo, A. (2006). The association between co-authorship network structures and successful academic publishing among higher education scholars.
Saxenian, A. (1994). *Regional advantage: Culture and competition in Silicon Valley and Route 128*. Cambridge: Harvard University Press.
Schiffauerova, A., & Beaudry, C. (2012). Who owns the intellectual property and where? The case of Canadian biotechnology. *International Journal of Biotechnology, 12*(3), 147–169.
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.
Schrader, S. (1991). Informal technology transfer between firms: Cooperation through information trading. *Research Policy, 20*, 153–170.
Taylor, M. S., Locke, E. A., Lee, C., & Gist, M. E. (1984). Type A behavior and faculty research productivity: What are the mechanisms? *Organizational Behavior and Human Performance, 34*(3), 402–418.
Tong, X., & Frame, J. (1994). Measuring national technological performance with patent claims data. *Research Policy, 23*, 133–141.
Toutkoushian, R. K., Porter, S. R., Danielson, C., & Hollis, P. R. (2003). Using publications counts to measure an institution's research productivity. *Research in Higher Education, 44*(2), 121–148.
Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32*, 425–443.
Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem! *American Journal of Sociology, 111*(2), 447–504.
van Zeebroeck, N., & van Pottelsberghe de la Potterie, B. (2006). Filing strategies and patent value.
von Hippel, E. (1987). Cooperation between rivals: Informal know-how trading. *Research Policy, 16*, 291–302.
Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A rexible data-analytic strategy. *Journal of Counseling Psychology, 34*(4), 372–382.

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge, ENG and New York: Cambridge University Press.

Watts, D. (1999). Networks, dynamics, and the small-world phenomenon 1. *The American Journal of Sociology, 105*(2), 493–527.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393*(6684), 440–442.

Yook, S. H., Jeong, H., Barabási, A. L., & Tu, Y. (2001). Weighted evolving networks. *Physical Review Letters, 86*(25), 5835–5838.

Appendix B. Appendices for Section 5.2.3

Table B.1. 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					
$ln\_avgFund3_{i-1}$	0.0341	1.0000				
$noArt3_{i-1}$	0.0314	0.4229	1.0000			
$ln\_avgCit3_{i-1}$	0.0552	0.1207	0.0742	1.0000		
ln_avgIf3_{i-1}	0.0764	0.1081	0.0481	0.4014	1.0000	
$careerAge_i$	0.0039	0.3201	0.2940	0.2047	0.0228	1.0000

Table B.2. 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					
$ln\_avgFund3_{i-1}$	0.0338	1.0000				
$noArt3_{i-1}$	0.0199	0.4229	1.0000			
$ln\_avgCit3_{i-1}$	0.0481	0.1207	0.0742	1.0000		
ln_avgIf3_{i-1}	0.0683	0.1081	0.0481	0.4014	1.0000	
$careerAge_i$	0.0056	0.3201	0.2940	0.2047	0.0228	1.0000

Table B.3. 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						
$ln\_avgFund3_{i-1}$	0.1467	1.0000					
$noArt3_{i-1}$	0.2559	0.4403	1.0000				
$ln\_avgCit3_{i-1}$	0.1031	0.1206	0.0899	1.0000			
ln_avgIf3_{i-1}	0.0386	0.1197	0.0547	0.4037	1.0000		
dc_i	0.0394	0.0459	0.0522	0.0586	0.1114	1.0000	
$careerAge_i$	0.0930	0.3406	0.3062	0.2361	0.0251	-0.0193	1.0000

Table B.4. Correlation matrix, clustering coefficient (cc) model

Variable	cc_i	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$dc*10\text{^}4_i$	$careerAge_i$
cc_i	1.0000						
$ln\_avgFund3_{i-1}$	-0.0982	1.0000					
$noArt3_{i-1}$	-0.2018	0.4403	1.0000				
$ln\_avgCit3_{i-1}$	0.0746	0.1206	0.0899	1.0000			
ln_avgIf3_{i-1}	0.0496	0.1197	0.0547	0.4037	1.0000		
$dc*10^4_i$	0.0571	0.0459	0.0522	0.0586	0.1114	1.0000	
$careerAge_i$	-0.0686	0.3406	0.3062	0.2361	0.0251	-0.0193	1.0000

Table B.5. 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						
$ln\_avgFund3_{i-1}$	0.0088	1.0000					
$noArt3_{i-1}$	0.0353	0.4403	1.0000				
$ln\_avgCit3_{i-1}$	0.0412	0.1206	0.0899	1.0000			
ln_avgIf3_{i-1}	0.0604	0.1197	0.0547	0.4037	1.0000		
dc_i	0.4916	0.0459	0.0522	0.0586	0.1114	1.0000	
$careerAge_i$	-0.0059	0.3406	0.3062	0.2361	0.0251	-0.0193	1.0000

Table B.6. Correlation matrix, closeness centrality (cl) model

Variable	cl_i	$ln\_avgFund3_{i-1}$	$noArt3_{i-1}$	$ln\_avgCit3_{i-1}$	ln_avgIf3_{i-1}	$dc*10^2{}_i$	$careerAge_i$
cl_i	1.0000						
$ln\_avgFund3_{i-1}$	0.0821	1.0000					
$noArt3_{i-1}$	0.0806	0.4656	1.0000				
$ln\_avgCit3_{i-1}$	0.0490	0.0936	0.0873	1.0000			
ln_avgIf3_{i-1}	0.1574	0.1013	0.0501	0.4421	1.0000		
$dc*10^2{}_i$	0.3845	0.0329	0.0318	0.0496	0.1619	1.0000	
$careerAge_i$	-0.144	0.3808	0.3175	0.0768	-0.0285	-0.0493	1.0000

Table B.7. Regression result, distinct team size model

teamSize	***Coef.***	***Std. Err.***	***t***	***P>|t|***	***[95% Conf. Interval]***	
ln_avgFund3	.8354446***	.1327287	6.29	0.000	.575296	1.095593
$noArt3_{i-1}$	.1410797*	.0762243	1.85	0.064	-.0083201	.2904795
$ln\_avgCit3_{i-1}$	.6037981***	.1219042	4.95	0.000	.3648656	.8427306
ln_avgIf3_{i-1}	2.460513***	.1986399	12.39	0.000	2.071179	2.849848
careerAge	-.4348039***	.137	-3.17	0.002	-.7033244	-.1662834
$careerAge^2$	.0235044***	.0084145	2.79	0.005	.0070119	.0399968
Affiliations dummy variable						
dAcademia	-4.633412***	.7062831	-6.56	0.000	-6.017729	-3.249095
_cons	-.7815731	1.450022	-0.54	0.590	-3.623621	2.060474

Notes: * p<0.10, ** p<0.05, *** p<0.01, number of observations: 60,907

Appendix C. Appendices for Section 5.3.1

Lift Charts

Lift chart is a tool to measure the performance of a model in classifying the data. As it can be seen in Figure C.1., in Task A around 80% of the data has been classified by the proposed model with the confidence of higher than 83%. For Task B (Figure C.2.), the model was succeeded to classify about 90% of the data with higher than 96% of the confidence level. According to Figure C.3., the proposed model has classified more than 90% of the data with the confidence of 99% and higher that shows the reliability of the proposed model in classifying the data in Task C. In addition, according to the curves it can be seen that the model reacts relatively fast to the data and it does increasingly better as it gets more data.

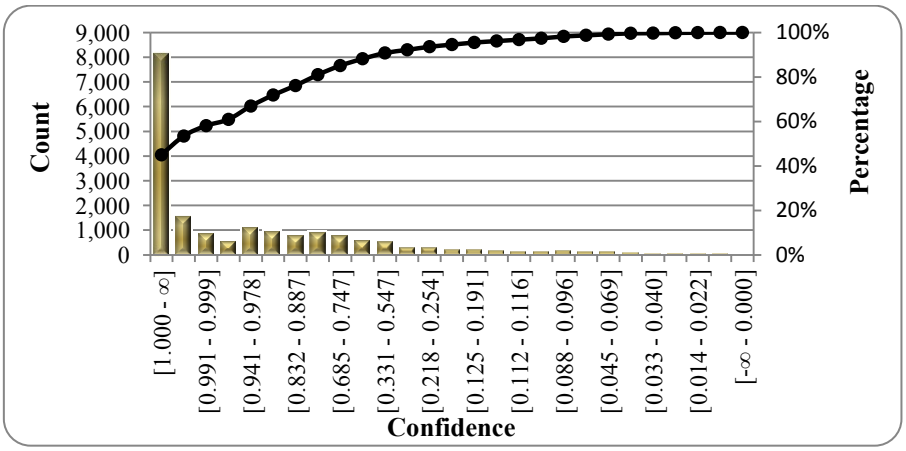


Figure C.1. PCF lift chart, Task A

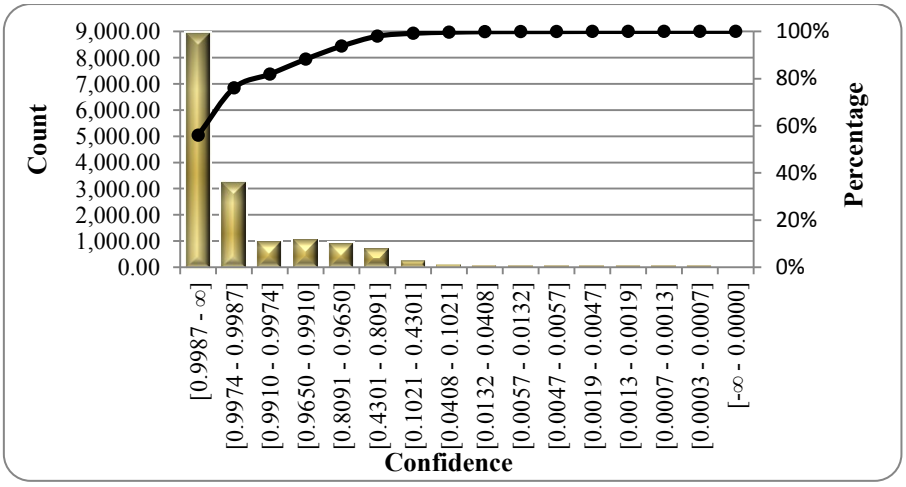


Figure C.2. PCF lift chart, Task B

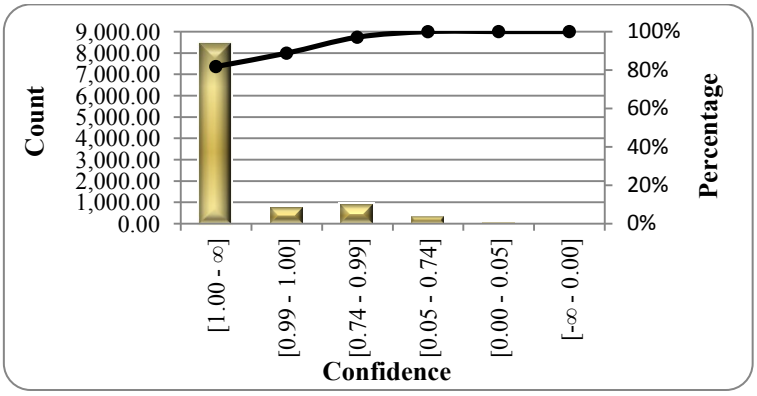


Figure C.3. PCF lift chart, Task C

Sample of Prediction Results

Sample of the predictions for both tasks (Task1 and Task 2) are presented in this part. The definition of the attributes was listed in Table 30. The real number of articles is shown in the *noArt* column that was not fed into the framework. Based on the other defined attributes the framework has predicted the number of publications that is highlighted in dark grey in the tables.

Table C.1. Sample of prediction results, Task 1

Predicted no articles	noArt	sum Fund3	avg If3	avg Cit3	team Size	btwn3	clust3	deg3	eigen3	career Age	discip	noArt3
0.361	0	0.041	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.737	2	0
1.102	0	0.013	0.279	0.028	0.000	0.000	1.000	0.005	0.000	0.632	3	1
3.865	7	0.044	0.054	0.005	0.001	0.059	0.125	0.027	0.000	0.737	1	13
1.103	0	0.010	0.068	0.083	0.000	0.000	1.000	0.007	0.000	0.737	3	1
1.206	1	0.072	0.132	0.020	0.002	0.016	0.409	0.020	0.000	0.526	0	6
6.703	4	0.167	0.246	0.080	0.002	0.055	0.158	0.039	0.000	0.737	1	26
1.030	4	0.032	0.115	0.017	0.001	0.018	0.455	0.018	0.000	0.737	0	6
4.120	3	0.061	0.136	0.041	0.002	0.185	0.109	0.134	0.000	0.737	1	15
0.000	0	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.263	0	0
5.047	3	0.137	0.141	0.041	0.001	0.133	0.163	0.050	0.000	0.684	0	15

Table C.2. Sample of prediction results, Task 2

Predicted Fund	sum Fund	sum Fund3	avg If3	avg Cit3	team Size	btwn3	clust3	deg3	eigen3	career Age	discip	noArt3
$414,936	$53,515	0.205	0.189	0.092	0.002	0.008	0.222	0.009	0.000	0.579	1	0.096
$70,832	$69,786	0.023	0.141	0.010	0.002	0.000	0.600	0.005	0.000	0.474	1	0.019
$60,750	$51,880	0.011	0.132	0.019	0.002	0.000	0.444	0.008	0.000	0.737	2	0.019
$183,301	$239,331	0.072	0.150	0.042	0.001	0.016	0.409	0.011	0.000	0.526	0	0.058
$78,938	$49,918	0.023	0.178	0.019	0.000	0.001	0.500	0.004	0.000	0.684	1	0.019
$158,689	$159,600	0.073	0.140	0.010	0.001	0.007	0.400	0.005	0.000	0.526	1	0.019
$131,313	$114,421	0.042	0.096	0.070	0.002	0.048	0.257	0.014	0.000	0.737	0	0.077
$117,806	$88,280	0.043	0.101	0.029	0.001	0.001	0.333	0.004	0.000	0.737	0	0.019
$85,018	$58,800	0.022	0.080	0.019	0.001	0.000	0.000	0.001	0.000	0.368	0	0.010
$74,211	$106,750	0.017	0.051	0.074	0.001	0.000	1.000	0.004	0.000	0.105	0	0.019