HEC MONTRÉAL

Affiliée à l'Université de Montréal

Analysis of the Evolution of the Scientific Collaboration among the HEC Montréal Professors

par

Behnoosh Saboonchi

HEC Montréal, Science de la gestion (Intelligence d'affaires)

Thèse présentée à la faculté des études supérieures et postdoctorales en vue de l'obtention du grade de Mâitrise és sciences (M. Sc.)

Mai 2014

HEC Montréal
Affiliée à l'Université de Montréal

Cette thèse intitulée :

Analysis of the Evolution of the Scientific Collaboration among the HEC Montréal Professors

présentée par
Behnoosh SABOONCHI

a été évaluée par un jury composé des personnes suivantes :

Professeur Pierre HANSEN
Directeur de recherche

Professeur Gilles CAPOROSSI
Codirecteur de recherche

Professeur Allain JOLY
Membre du jury

Professeur Sylvain PERRON
Membre du jury

Résumé

Dans ce mémoire, nous avons étudié le réseau de collaboration des professeurs de HEC Montréal à l'aide de la base de données fournie par la Direction de la recherche de HEC Montréal sur leurs publications conjointes.

En utilisant la théorie des réseaux, nous avons établi une présentation en graphe de la collaboration de publication et étudié son évolution au fil du temps en utilisant des statistiques descriptives, des méthodes d'analyse des réseaux sociaux et d'analyse dynamique de réseau.

Au cours des dernières années, les réseaux de collaboration reçoivent de plus en plus d'attention et sont le sujet de beaucoup de recherches. Il y a des raisons derrière cela : d'abord avec l'aide des nouvelles technologies, ils sont devenus faciles à documenter et à suivre, et ensuite, ils sont de bons exemples de réseaux sociaux. Si deux auteurs ont publié un article ensemble, il y a de fortes chances qu'ils se connaissent en personne, donc ces réseaux représenteraient leur interaction sociale [41]. Des recherches récentes démontrent que les réseaux de collaboration scientifique influencent les pratiques scientifiques [32].

Ces raisons nous ont motivé à mener cette étude et mettre en œuvre les méthodes d'analyse des réseaux sociaux sur le réseau des professeurs de HEC Montréal. Ceci est la première analyse qui est menée sur le dossier de publications au sein de l'école. En outre, la plupart des réseaux sociaux étudiés dans la littérature sont statiques ce qui ne tiennent pas compte de l'évolution et l'émergence du réseau, tandis que dans ce travail, nous avons également utilisé les méthodes d'analyse dynamique de réseau pour étudier les changements et l'évolution du réseau de publications au fil du temps (i.e. de 2000 jusqu'à 2011).

Les résultats de l'analyse de notre réseau étudié montrent certaines similarités avec d'autres réseaux de collaboration étudiés dans la littérature, tels que, le phénomène de petit monde, la présence du coefficient de clustering, la structure de la communauté et l'attachement préférentiel. Sans parler de certains motifs de collaboration entre les différents départements et les communautés sont exposés et présentés dans ce travail.

La contribution de ce travail est de combiner diverses techniques d'analyse de réseaux sociaux qui sont fragmentées dans la littérature, pour analyser le réseau de collaboration académique.

Summary

In this work we studied the collaboration network of the HEC Montréal professors using the database provided by the Research Direction of HEC Montréal (Direction de la recherche de HEC Montréal) on their joint publications.

Using the network science, we established a graph network presentation of the publication collaboration and studied its evolution over time using descriptive statistics, social network analysis and dynamic network analysis methods.

In recent years the collaboration networks are getting more attention and are the focus of much research. There are some underlying reasons for that: first with the help of new technologies they have become easy to map, document and track, and second they are good examples of social networks because if two authors have published a paper together, there is a high chance that they know each other in person so that such networks would represent their social interaction [41], and finally recent works show that scientific collaboration networks affect scientific practices [32].

These reasons motivated us to conduct this study and implement the social network analysis methods on the joint publications network of the HEC Montréal professors which is the first time such analysis is conducted on the publication records within the school. Besides, most of the social networks studied in the literature are static which do not take into account the evolution and the emergence of the network, whereas in this work we also used the dynamic network analysis methods to study the changes and evolution of the publications network over time (i.e. from 2000 to 2011).

Our analysis show some similarities with other collaboration networks studied in the literature, such as the small-world phenomenon, the presence of the clustering coefficient, the

community structure and the preferential attachment. Besides, some patterns of collaborations among different departments and communities are uncovered and presented in this work. Finally, the contribution of this work is that it combines various fragmented social network analysis techniques available in the literature to analyze the academic collaboration network.

KeyWords : Collaboration networks, Social Network analysis, Dynamic Networks, Preferential attachment, Community detection

Table of Contents

List of Tables

List of Figures

Aknowledgements

This thesis would have been impossible without the guidance of my thesis supervisor Professor Pierre Hansen. Thanks for accepting to supervise this work, I could not have asked for more supportive and knowledgeable role model. Your constant encouragement, support, and invaluable suggestions made this work successful.

Gilles Caporossi, my dear co-supervisor! I feel very privileged to have worked with you. I have always admired your great personality and energy. I would also like to thank professor Allain Joly for accepting to evaluate and improve this work.

I would like to thank my friends for their support, care and encouragement which helped me overcome my stress and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me. Cheers to Lysiane, Hakim, Mylène and William. Thanks for being there for me, the long days of studying would not have been possible without you.

I'd also like to thank the great people at GERAD and HEC Montréal for creating this welcoming environment, especially Carole Dufour.

Finally, I am dedicating this thesis to my lovely parents, my sister and Nicolas for their unconditional love and support throughout my life!

General Introduction

The Network science has been the focus of much research during the recent years owing to its capability to analyze the behaviour of complex systems. This interdisciplinary science helps develop practical techniques to better understand both natural and man-made networks [11] and is capable of studying even larger networks with the presence of high speed computers and the internet to store and track data.

Social network is a type of network which became much popular after the rise of Facebook. While social networks hold the fundamental characteristics of general networks, they have some special features. Newman and Park [41] discussed that the first difference would be the "assortative mixing" [41] of nodes in social networks which is the tendency of the nodes with similar degrees to connect with each other and the second difference is the presence of clustering coefficient in social networks where it is more likely for two nodes that have a node in common to connect with each other.

The scientific collaboration of the professors which is the focus of this thesis is a form of social network since it possesses the special characteristics of the social networks. Newman and Park [41] believe that the collaboration networks are good examples of social networks because if two authors have published a paper together, there is a high chance that they know each other in person so that such networks would represent their social interaction.

Most of the social networks studied in the literature are static which do not take into account the evolution and the emergence of the network. Dynamic Network Analysis (DNA) on the other hand, considers the dynamics and evolution of the network over time [6]. In this work we use the publication database of HEC Montréal professors between 2000 and 2011, and since we study its changes and evolution over time, it forms a dynamic network.

The first chapter presents the literature review and starts with the network science and its history. It then continues with a description of various kinds of social networks and ends by presenting the similar works to this study that have already been studied in the literature.

The second chapter studies the basic statistics of the network of HEC Montréal's publications between 2000 and 2011 in order to better understand and visualize how this network looks like, including the statistics on the number of papers, authors, degree distributions, etc. The results are first presented for each individual year statically and then the results are presented cumulatively.

The third chapter addresses the community detection problem in networks; the community structure is known to be common in many real networks. Finding the communities in the collaboration network of HEC Montréal will help understand what the profile and the background of the professors belonging to one community are and would also explain what common properties would bring the professors together in a community.

The fourth and last chapter studies the preferential attachment, i.e. the assumption that in scale-free networks the higher degree nodes acquire new nodes faster than lower degree ones [5, 16, 26]. This phenomenon is studied at both the nodes and the communities levels. At the level of nodes the following aspects have been studied: the behaviour of new authors entering the network for the first time, the effects of repeated collaborations, the department each professor belongs to and finally the number of common neighbours on the decision of co-authorship. And at the level of communities same analyses are conducted and it is studied what would be the size of the target communities that the nodes not belonging to any community would like to attach to.

The results show that the studied network forms a small-world, and the clustering coefficient, the community structure and the preferential attachment are present in our collaboration network. The dynamical analysis suggests that the average degree increases over time, and the average path length, average clustering coefficient and the relative size of the largest component increase only at the beginning of the studied period and tend to have a more stable value in the last years where the network has been shaped.

Finally, with the study of the preferential attachment, it is known that in our studied network, new nodes entering the network for the first time are more likely to attach to the

nodes having a bigger degree and also are more likely to attach to bigger communities. The results also show that a pair of nodes having three common neighbours are about twice more likely to collaborate compared to a pair of nodes with zero common neighbours. Finally, it is known that there is more tendency among professors in the same department to collaborate compared to the ones coming from different departments. Also, the contribution of this work is that it combines various fragmented social network analysis techniques available in the literature to analyze the academic collaboration network.

Chapter 1

Literature Review

1.1 Network and Network Science

In recent years the network and network science have received much attention within "biological, social, technological, and information networks" [21]. One of the underlying reasons is the capability of such models to represent and study different complex systems. In the following subsections these aspects will be discussed in details.

1.1.1 Network

The history of graph (network) dates back to 1736 when Leonhard Euler tried to solve the "Seven Bridges of Königsberg" problem and then proved that there is no solution for it [1, 19]. For him this problem belonged to a class of problems that Leibniz had previously called "geometry of position" where only the "determination of positions and its properties" were important and not the distances among the nodes [1, 19].

This was only the beginning of the graph theory; as Biggs, Lloyd and Wilson state their opinions on the above paper by Euler in their Graph theory book [9] : "The origins of graph theory are humble, even frivolous. Whereas many branches of mathematics were motivated by fundamental problems of calculation, motion, and measurement, the problems which led to the development of graph theory were often little more than puzzles, designed to test the ingenuity rather than to stimulate the imagination. But despite the apparent triviality

of such puzzles, they captured the interest of mathematicians, with the result that graph theory has become a subject rich in theoretical results of a surprising variety and depth."

1.1.2 Network Science

Network science is an interdisciplinary science that studies networks and "aims to develop theoretical and practical approaches and techniques to increase our understanding of natural and man made networks" [11].

The network science got popular at the beginning of the 21st century. Albert-Làszló Barabàsi in his "Network Science" [3] book believes that this popularity could be followed over time by the citation pattern of two of the most classic papers on the subject: 1) the paper on the introduction of random networks in 1959 by Paul Erdös and Alfréd Rényi [17] and 2) the most cited social network paper by Mark Granovetter in 1973 (Cited 25198 times as of August 23rd 2013 [24]). As seen in figure 1.1 [3] these two reference papers received exponentially more amount of attention after 2000.

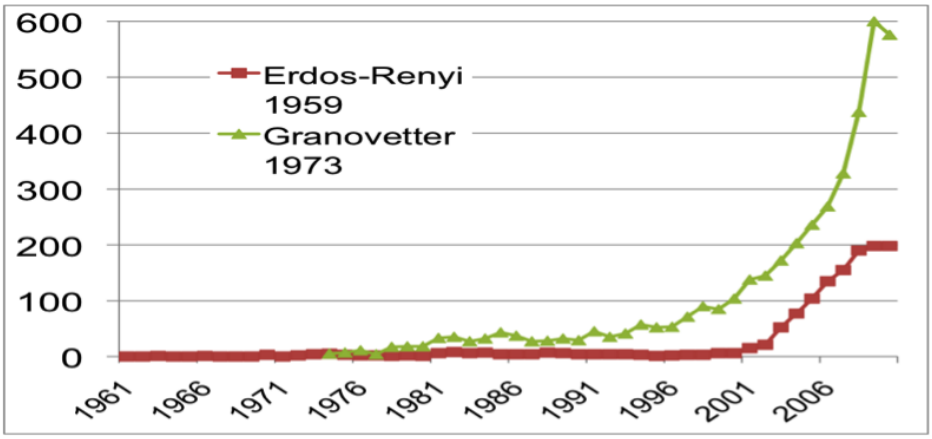


Figure 1.1: The Citation Pattern of Two Classic Network Science Reference Papers

Considering the fact that the network theory dates back to 1736, then why does this science get more attention at the beginning of the 21st century? Firstly, the presence of high speed computers and the internet made the storing and tracking of the network maps a whole lot easier [3]. In order to analyze huge and complex networks such as the citation network of scientific publications one would need resources to store these publications and

track their evolution overtime, which sounded impossible back in the 1950s when Erdös and Rényi [17] first introduced their random network model. Secondly, Barabási [3] considers the "universality of network characteristic" as an aspect of network science that attracts more attention to this domain of research. He argues that although there are a lot of differences in various networks in nature, size, evolution, history etc., their architectures are rather similar which allows for the use of a common set of mathematical tools to explore them [3].

1.2 Social Network Analysis (SNA)

Scientists in various fields such as sociology, biology, chemistry etc., use network sciences to map and analyze their networks. Network science had an important role in even fighting against terrorism. As an example in 2003 the US military constructed the social network of Saddam Hussein's entourage that eventually led them to his hiding place [3].

Social network is one of the most known applications of network science and has gained a large popularity after the rise of Facebook. Newman and Park [41] discussed how social networks are different from non social networks. The first difference would be the "assortative mixing" [41] of nodes in social networks, which is the tendency of the nodes with similar degrees to connect with each other. This characteristic is not present in non social networks where degrees are negatively correlated, i.e., "Disassortative mixing" [41]. The second difference is the presence of clustering coefficient in social networks where it is more likely for two nodes that have a node in common to connect with each other. Within the same paper they have shown that these two differences are due to the division of the nodes into communities in social networks [41].

1.3 Collaboration Networks

The scientific collaboration network where two scientists are considered as connected when they publish at least one paper together is a type of social network that is the main focus in this work. Newman and Park [41] believe that the collaboration networks are good examples of social networks because if two authors have published a paper together, there is a high chance that they know each other in person so that such networks would represent their social interaction. Also "recent work in the sociology of knowledge demonstrates a direct linkage between social interaction patterns and the structure of ideas, suggesting that

scientific collaboration networks affect scientific practice" [32]. Besides, these networks are very well documented, are easy to map and track with the help of new technologies and also there is a clear definition of acquaintance between the nodes (scientists), whether they have published together or not.

The publication record databases have been used before to study the *co-citation* "(i.e., connections between authors established via the citation of their works in the same literature patterns)" and co-authorship patterns [36]. Yet, Newman used the same databases to also study the *collaboration of authors* for the first time in 2000 [36]. He states that the closest research conducted on the same networks would be "the concept of Erdös number" [36] which is famous amongst the mathematicians' communities. The Erdös number measures the collaborative distance of scientists from the great Hungarian mathematician Paul Erdös; this number would be equal to one for the direct collaborators of Erdös, two for the collaborators of Erdös collaborators and so on and so forth [36].

The very first paper published using the collaboration networks by Newman [36] covers the basic analysis of these networks. He used three databases on three different subjects: computer science, biomedical research and physics. He realized some similarities among these three databases. For instance, in all of them communities form a *small world*, i.e., each pair of the nodes are connected to each other through "a short chain of acquaintances" [49], which in these databases the author found out that there are about five to six steps to get from one scientist in the community to another. Another similar characteristic is the presence of clustering which is a characteristic of social networks that distinguish them from other types of networks. He then presents some results on the number of each author's collaborators and the number of papers published by each of them. It is also demonstrated that both numbers (each author's collaborators and number of published papers) follow the *power-law distribution*, i.e. the fact that the nodes in such networks are not connected to each other randomly or in other words the professors do not collaborate at random and there are parameters that affect the decision of co-authoring. This characteristic which is called *preferential attachment*, will be discussed in detail in Chapter 4.

After a general analysis of the three databases mentioned above, Newman published another paper [35] using the very same databases with more details and by presenting

theoretical measurements: 1) *Shortest paths:* the smallest distance between two nodes is called the shortest path. In collaboration networks would be the smallest distance, if any, between two scientists. The *average* of shortest paths between pairs of nodes is also of interest, called the *average distance*, 2) *Betweenness:* with this measure one could identify the most influential scientists in the network, or in a more scientific way, the betweenness of a node is "the total number of shortest paths between pairs of actors that pass through this node." [35], 3) *Funneling effect:* Strogatz [35] brought up the question of whether all of your collaborators have the same role to connect you to the rest of the network or rather one author has such influential and important collaborators that all of the shortest paths from that author to the others would pass through them? The answer would be that the shortest paths for the majority of scientists would pass only through one or two of their collaborators [35]. In other words collaborating with well connected and influential authors in each field would assure one's connection to a "large part of the collaboration network" [35], and finally, 4) *Closeness:* this measure is the average of distances of each node with other nodes and the lower the closeness for one scientist the more central that scientist is.

An interesting aspect of this paper is the presentation of a new metric to measure the strength of collaboration amongst scientists. Newman [35] argues that the usual and common metrics are not adequate to well explain the strength of the collaboration between scientists, as a result, he came up with a new metric which is called the *measure of collaboration* which is based on "the number of papers co-authored by pairs of scientists, and the number of other scientists with whom they co-authored those papers" [35]. In order to calculate this measure he constructed the weighted network by taking into account the number of papers the pairs of authors have co-authored and the number of other authors they have co-authored with. This measure is believed to better capture the collaborative ties among scientists. Once the weighted network is constructed all the metrics mentioned above would be recalculated so that they would have a better representation of the network structure.

There are several similar papers on the collaboration and social networks in general with the same focuses: analyzing some statistics to better understand the situation such as the number of nodes, the distances between the nodes and investigating the small-world property, etc. There are other papers in the literature similar to the three papers discussed

above by Newman which study more or less the same concepts and use the same approaches to analyze social networks [2, 25, 31, 34, 48].

Another interesting finding in Newman's work is that depending on the domain of the co-authorship network the collaboration pattern might vary. As a result, to elaborate more on those differences he published another work in 2004 on the co-authorship networks and patterns of scientific collaboration within different domains [38]. Working with the same three databases as before, he finds out that the number of co-authors for each scientist and the number of authors per paper vary depending on the field of study, as an example biological scientists have more co-authors than mathematicians or physics scientists. The author believes that this is the reflection of "labor intensive, predominantly experimental direction of current biology" [38]. The other difference observed among these databases is the clustering coefficient that is higher for physics and lower for biology [38], meaning that having a mutual collaborator in biology would less likely result in a co-authorship compared to the same situation in physics. He argues that the answer to the question why authors in different fields show different clustering coefficients is not evident and not easy to answer, though he states that part of the issue could be explained by the fact that the papers with two or three authors would result in a bigger coefficient and the other part could be explained by different "sociological or organizational effects" within various domains [38].

1.4 Dynamic Network Analysis (DNA)

Social networks have been widely studied in the literature, yet most of them are static which does not take into account the evolution and the emergence of the network. Dynamic Network Analysis (DNA) on the other hand, considers the dynamics and evolution of the network over time [6].

Carley [13] compares dynamic network analysis with quantum mechanics and name some similarities between them. In DNA similar to quantum mechanics, "the relations are probabilistic, the measurement of a node changes its properties" and also "movement in one part of the system propagates through the system" [13]. It is obvious that with this approach the nodes and edges are subject to change over time. In our case regarding the collaboration network, the network is constantly expanding by the addition of the new nodes (authors)

and also the addition of the new edges (new collaborations between existing authors or a collaboration with newly added authors). It is also possible for the edges and the nodes to disappear from the network.

Chapter 2

HEC Montréal's publications database characteristics

In this chapter we introduce the database used in this work which is the network of the HEC Montréal's professors's joint publications between 2000 and 2011. This database is provided by the Research Direction of HEC Montréal (Direction de la recherche de HEC Montréal) which includes the publications of not only schools' professors but also other authors such as their students or external collaborators.

We would first study the basic statistics of this database to better understand and visualize how the network looks like, including the statistics on the number of papers, authors, degree distributions, etc. The results are first presented for each individual year statically, and then the results are presented cumulatively (by adding one additional year at a time) to better visualize the evolution and growth of the network during these 12 years.

2.1 Data Preparation

As mentioned above, the database contains also the students and external authors as well as HEC Montréal's professors. As a result, in order to separate the professors from other authors, a complete list of professors, their titles and departments have also been provided by the Human Resources of the school. Thereafter, all of the authors that do not belong to this list have been removed and the network has been constructed with only professors as the nodes and their publications as the edges whenever they published a paper together.

The reason why all the external authors have been removed is that we intend to concentrate our study only on the professors of HEC Montréal and their interactions.

2.2 The Static Analysis of the Co-authorship Network

We first analyze the complete 12-year network by including all the data from 2000 to the end of 2011. Such network has been constructed by putting together all the nodes (the professors) and the edges (their publications) at the end of the studied period, i.e. 2000 to 2011.

2.2.1 Basic Statistics

In Table 2.I the basic statistics have been presented for each year separately and also for the whole network at the end of 2011.

Table 2.I : Basic Statistics for HEC Montréal Publications Network

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	All Years
Number of Papers	62	147	172	178	186	236	204	216	202	**271**	241	248	2363
Number of Authors	46	78	90	95	109	128	126	122	113	**146**	122	125	271
Giant Component	4	4	5	6	5	7	9	11	13	17	7	7	185
Average Distance	1.375	1.375	1.472	1.623	1.382	1.608	1.964	1.984	2.415	2.734	1.774	1.633	5.831
Maximum Distance	3	3	3	4	3	4	5	4	7	7	6	4	15
Clustering Coefficient	0	0	0	0.185	0.167	0.137	0.197	0.084	0.198	0.173	0.203	0	0.153
Graph Density	0.011	0.007	0.005	0.007	0.006	0.005	0.006	0.007	0.008	0.006	0.006	0.006	0.008

The highest number of papers published and also the highest number of authors in one year happens in 2009, which is 146 authors who published 271 papers.

A *component* in a network is a set of connected nodes in a way that there is a path between any given node within that component to any other member [38]. The component with the largest number of nodes in the whole network is called the *giant component*. As seen in this table, the year 2009 possesses also the largest giant component. The overall size of the giant component for the whole 12-year network is 185 which accounts for 68.26% of total nodes (total 271 professors) which represents a solid interconnectedness among the professors in HEC Montréal.

The average distance which is the average of shortest paths between pairs of nodes for the overall network in 2011 is equal to 5.831, meaning that on average there are 5.831 steps

between every two professors (if connected at all). The maximum distance, or the network diameter shows that the maximum connecting steps between any two professors (nodes) is equal to 15.

The clustering coefficient which is the probability that two neighbours of a node collaborate together [6] is equal to 0.153 for the overall 12-year network.

The graph density is the ratio of all the edges present in the network over all the possible edges among pairs of nodes. A complete graph has a density equal to one, in our case for the overall 12-year network this ratio is equal to 0.008 which indicates that the HEC Montréal publication network is not a dense one.

2.2.2 Distribution of the Professors Titles

In Table 2.II the distribution of the professors' titles for the whole network from 2000 to 2011 is presented. It should be noted that the titles used in this table correspond to the latest title of each professor and the changing of the professors' ranks during the previous 12 years has not been captured.

The first column represents the number of professors that have published at least one paper between 2000 to 2011 within each rank. The second column is the total number of professors working at HEC Montréal at each rank regardless of their publication status, and finally the last column shows the percentage of the publishing professors within each rank.

Table 2.II : Professors' Titles Distribution by the End of 2011

Professors' Titles	Count	Overall	%
Assistant Professor	125	157	79.62
Associate Professor	71	92	77.17
Full Professor	57	71	80.28
Guest Professor at the rank of an associate	10	16	62.5
Guest Professor	3	6	50
Visiting Professor	3	4	75
Visiting professor at the rank of a full professor	1	1	100
Visiting Professor at the rank of an associate	1	2	50
Assistant Researcher	0	1	0
Substitute Professor	0	1	0
Total	271	351	77.21

In the studied network around 91% of the professors are assistant, associate or full professors with almost similar percentage of members that have published at least one paper between 2000 to 2011.

2.2.3 Distribution of the Professors' Departments

In Table 2.III the distribution of the number of professors in terms of the latest department they belonged to is presented.

The first column represents the number of the publishing professors from each department, the second one is the total number of professors within each department at HEC Montréal regardless of their publication status, and finally the third column shows the percentage of the publishing professors within each department.

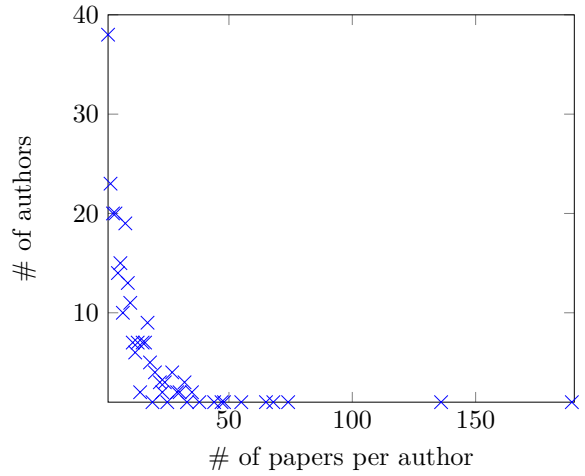
Table 2.III : Professors' Departments Distribution by the End of 2011

Department	Count	Overall	%
Department of Logistics and Operations Management	40	51	78.43
Department of Marketing	37	46	80.43
Department of Accounting Studies	34	49	69.39
Department of Finance	32	51	62.74
Institute of Applied Economics	30	41	73.17
Department of Management	27	30	90.00
Department of Management Sciences	24	27	88.89
Department of Human Resource Management	21	22	95.45
Department of Information Technologies	20	23	86.96
Department of International Business	6	11	54.54
Total	271	351	Avg. 79.66

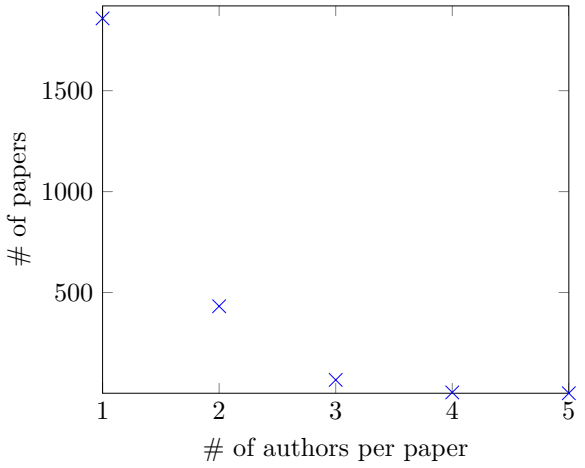
The average of the percentage of the publishing professors within each department is 79.66 which indicates that on average the majority of the professors in each department have had at least one publication between 2000 to 2011.

2.2.4 Number of Papers per Author and Number of Authors per Paper

In Figure 2.1 the distribution of the number of papers per author (a) and also the distribution of the number of authors per paper (b) are presented.



(a)



(b)

Figure 2.1: Distibution of (a) Number of Papers per Author and (b) Number of Authors per Paper

As seen in Figure 2.1 (a), a smaller percentage of authors contribute to the majority of the number of publications, i.e. around 10% of the professors account for 80% of the total publications between 2000 to 2011. The weighted average of the number of published papers per author is 11.28 which indicates that around 68% of the professors have below

average number of publications. At two extremes, there is an author that have published 149 articles, and 38 authors that have only published a single article.

In Figure 2.1 (b) it is observed that the majority of the papers have one or two authors, i.e. around 97% of the papers are published with less than three authors. The weighted average of the number of authors per paper is 1.25. At two extremes, there is a paper that has 5 authors, and 1858 papers that have only a single author.

It has to be noted that the studied database contains only HEC Montréal's professors and all the external authors have been removed. As a result, a single-author paper is not necessarily authored by only one person in first place, yet in this work we are only interested in the interactions among the professors of the school.

2.2.5 Degree Distribution

The degree of each node is the number of links connected to that node, i.e. in our network this degree is the number of collaborators of each professor. The probability distribution that a given node has k links is called the *degree distribution*. Barabási et al. [4] show that in large and complex networks, "independent of the system and the identity of its constituents", the degree distribution "decays" as power-law ($p(k) \sim k^{-\nu}$); indicating that large networks grow into a "scale-free state" [4], which could be explained by the preferential attachment [4]. This means that the nodes in such networks are not connected to each other randomly which in our case means the professors do not collaborate randomly and there are parameters that affect the decision of co-authoring.

Figure 2.2 illustrates the histogram of the distribution of the number of collaborators of each professor.

It is observed that the degree distribution of the network is fat-tailed meaning that the existence of a relatively small number of high degree nodes and a relatively large number of low degree nodes. As mentioned earlier, this behaviour could be explained by preferential attachment which is the assumption that in scale-free networks the higher degree nodes acquire new nodes faster than the lower degree nodes [5, 16, 26]. This aspect will be studied in detail in Chapter 4.

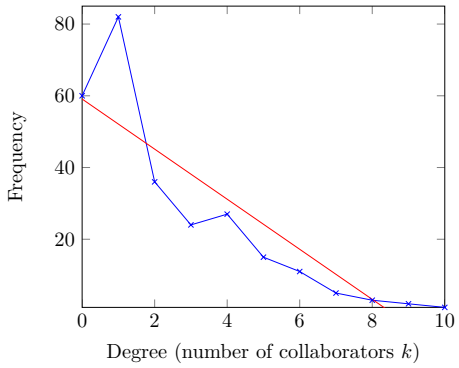


Figure 2.2: Histogram of the Degree Distribution with the Best Fitted Line

2.3 The Cumulative Analysis of the Co-authorship Network

In the previous sections we studied the network statically and in the following sections we would start to analyze it dynamically to better visualize and study its evolution over time. To do so, we would first construct the cumulative network starting from the very beginning with the articles published in 2000, and as the next step we would combine this network with the one in 2001 which would result in the combination of publications from 2000 to 2001, and we would continue the same process until the final cumulative network would contain the whole publications over the 12-year period.

2.3.1 Cumulative and Non-Cumulative Representation of the Number of Papers and Authors

Previously in Table 2.I we presented the number of papers published each year. Now Figure 2.3 (a) illustrates the same numbers in a graph and Figure 2.3 (b) visualizes their cumulative sum. It is observed that as time goes by, the number of publications per year increase and on the contrary the number of new authors added each year decrease which indicates the maturity of the collaboration among the HEC Montréal's professors.

In Figure 2.3 (c) the number of new authors added each year is presented where a new author is a professor that has not been present in the database up to a given date. This number is relatively decreasing over time, and finally in Figure 2.3 (d) the cumulative number of authors in each year is illustrated.

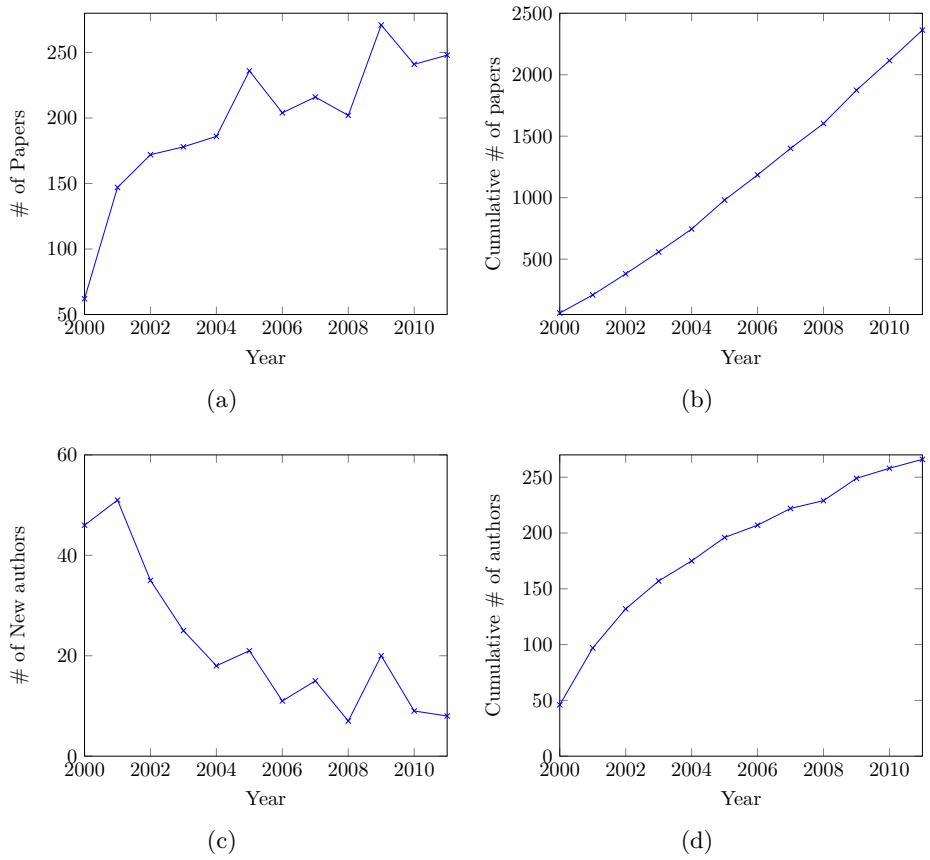


Figure 2.3: (a) Number of Papers Published Each Year. (b) Cumulative Number of Papers Published Between 2000-2011 (c) Number of New Authors Added Each Year to the Network. (d) Cumulative Number of Authors Between 2000-2011.

2.3.2 Average Path Length

The shortest path between two nodes is called the *path* between those nodes. The *average path* or the *average separation* is the average of the paths between each pair of nodes in the network. The shortest paths in our network are calculated using Gephi (an open-source network analysis software). This software uses the algorithm presented by Brandes in "a faster algorithm for betweenness centrality" [12]. In Figure 2.4 this average for the cumulative network is presented.

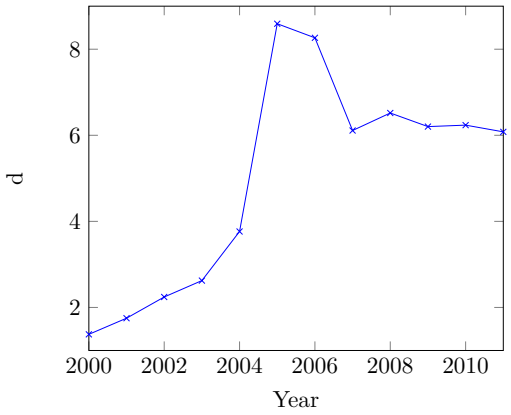


Figure 2.4: Average Path Length Based on Cumulative Database from 2000 up to any Year

The database studied in this research is not complete since it only contains the collaborations from 2000 which is for sure not the beginning of the collaboration of HEC Montréal professors. Since we are constructing the "already existing network" [6] after 2000, it would not be accurate to base our conclusions on the first years. Up until 2005 one observes that the network is being constructed and after that a more stable behaviour is formed in the network. Such phenomenon is observed for almost all of the metrics that we study in the database which is very similar to Barabási et al. observations in [6].

Going back to Figure 2.4, the average path length increases until 2005 and it starts to decrease thereafter. This shows that the nodes/professors are getting closer to each other due to more connections and collaborations overtime.

Another interesting point is that at the end of the studied period in 2011, the average path between the nodes is about six, meaning that on average any two professors, if connected at all, would be connected via six other authors. To conclude, the studied network forms a *small world* where nodes are connected together through "a short chain of acquaintances" [49].

2.3.3 Average Degree

The degree of a node is the number of links connected to that node, in this case is the number of *unique* collaborators that an author has. This means that the repeated collaborations among authors have not been double counted in the analysis as only the degree of each node

is taken into account and not their weighted degrees. The average degree of the nodes is presented cumulatively in Figure 2.5.

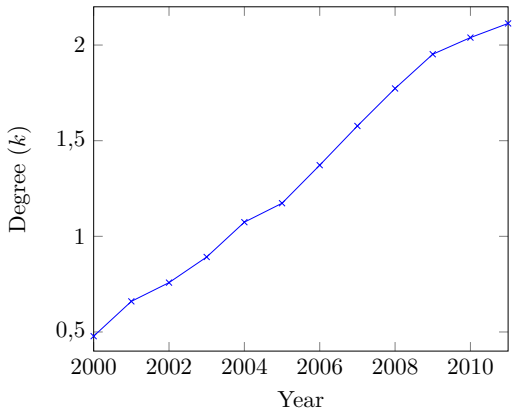


Figure 2.5: Average Degree of Authors Based on Cumulative Database from 2000 up to any Year

This average is increasing almost linearly over time and at the end of the period it reaches 2.11. This increase is the result of authors acquiring new collaborators, whether a connection with an author that has just joined the network or a new connection among already existing authors.

2.3.4 Average Clustering Coefficient

Clustering coefficient in simple words is the probability that the two neighbours of a node, collaborate together [6] and it answers the question "whether the collaborators of an author are connected together or not" [11]. This coefficient shows "how well connected are the neighbours of a node" [46] in the network. A more quantitative definition for this coefficient for a node would be the ratio between the number of edges between this node and its neighbours over all of the possible edges among them. The Clustering Coefficient is defined in the equation below where Δ is the number of triangles (clique of size three), and k_i is the degree of the node i [11]:

$$C = \frac{3 \times \Delta}{\sum_i \frac{k_i(k_i-1)}{2}}$$

This coefficient for the cumulative network up to year 2011 is presented in Figure 2.6.

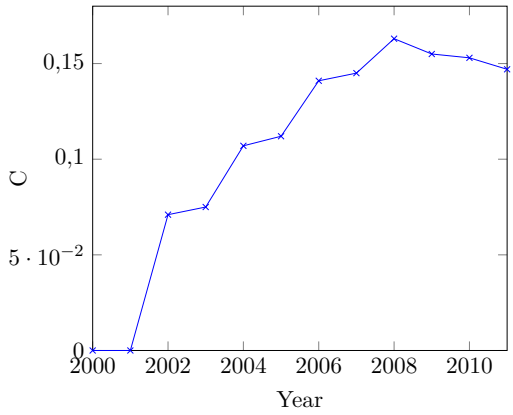


Figure 2.6: Average Clustering Coefficient Based on Cumulative Database from 2000 up to any Year

As seen in the figure above, the clustering coefficient is increasing fast until 2005 and after that it increases more steadily and ends in an almost constant clustering coefficient (~ 1.15) during the last years.

2.3.5 Relative Size of the Largest Component

A *component* is a set of connected nodes in a way that there is a path between any given node to any other member within that component [38]. The component with the largest number of nodes in the whole network is called the *giant component*. In our studied network the largest (giant) component includes 68% of the total network nodes, which indicates a solid interconnectedness among HEC Montréal professors.

In Figure 2.7 the ratio of the size of the largest component over the total number of the professors available in the network is presented cumulatively for each year.

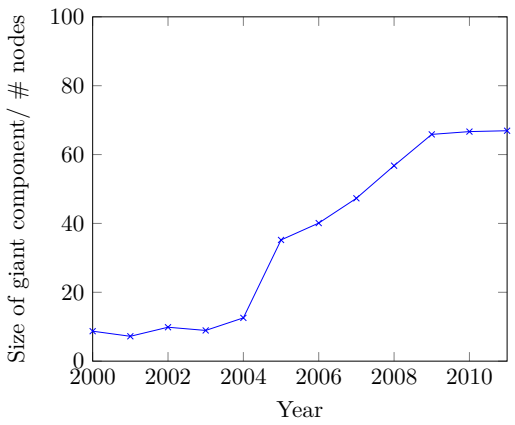


Figure 2.7: Relative Size of the Largest Component Based on Cumulative Database from 2000 up to any Year

During the first years, the largest component does not constitute a big proportion of the whole network. In the following years, as professors make connections by collaborating with authors coming from other communities, this component starts to grow. As we can observe from the above figure, around year 2005 this ratio starts to grow at a more rapid rate and it reaches an almost constant ratio ($\sim 68\%$) in the last years. As explained before this is also due to the fact that we are constructing an "already existing network" [6], i.e., it is possible that two authors have published an article before 2000 and that is not captured and as time goes by the network starts to taking shape. Therefore, during the last years the "basic alliances are almost uncovered" [6] which results in the constant value of the ratio.

2.4 The Evolution of The Biggest Components Over Time

By looking at the network statically in a single year we would observe a large number of isolated clusters. An underlying reason is that looking at the network in a single year neglects the collaborations that had occurred previously. Besides, some professors tend to publish papers individually or some would always work within the same group without making any connections with other communities [6].

Now, by looking at the network cumulatively/dynamically, one would observe that a giant component is developing and taking shape as years go by. Barabàsi et al. [6] mention that in most of research fields all of the professors, besides a small group of them that do not collaborate in groups, belong to the giant component and the network is "fully

connected" since the beginning. Newman [38] also mentions that "intellectual isolation from the mainstream of one's research area cannot often be a good thing", so authors tend to collaborate more in groups and exchange their knowledge.

To investigate more on how this giant component is taking shape in the studied network, the giant component in the cumulative database is presented in Figures 2.8 to 2.18 (in 2003 and 2004 the two and three most largest components are presented). Each colour represents a community (details on how communities are found is explained in Chapter 3), and the nodes are sized according to their betweenness centrality value, i.e., the bigger nodes have a bigger value of betweenness (more details on this value is discussed in Section 3.6).

As previously discussed in Section 2.3.5, the size of the giant component increases over time, which is also visually observed in the following figures. The giant component takes shape as the smaller components combine together. Once again in the year 2005 (Figure 2.12) we notice a shift in the evolution of the network and one is able to see the presence of the *one* giant component.

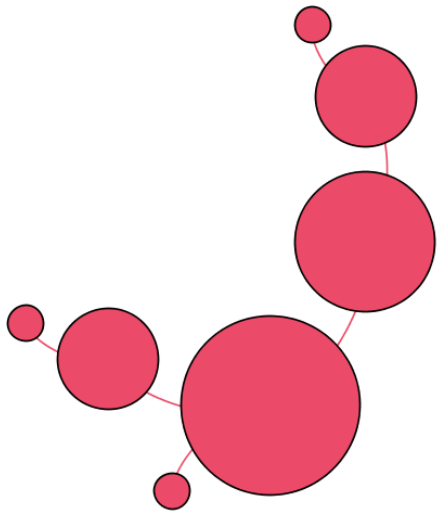


Figure 2.8: 2000-2001; the giant component in this period consists of 7 authors that accounts for $\sim 7\%$ of the whole network.

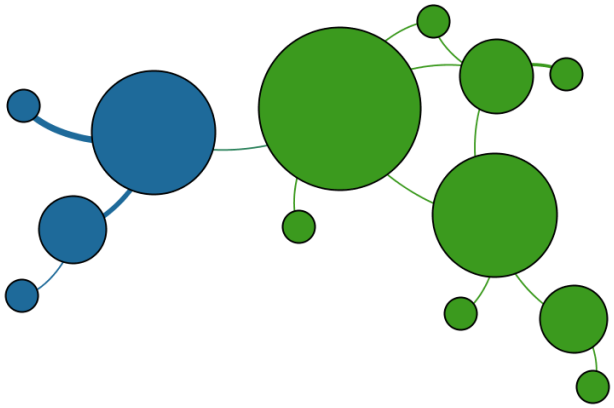


Figure 2.9: 2000-2002; the giant component in this period consists of 13 authors that accounts for $\sim 10\%$ of the whole network.

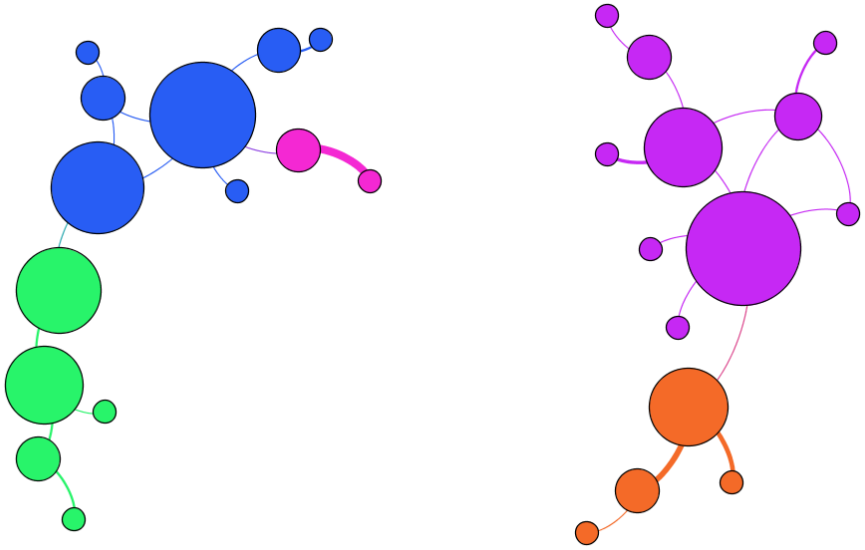


Figure 2.10: 2000-2003; in this period the two largest components are presented since they have the same size. They consist of 14 authors which accounts for $\sim$ 9% of the whole network.

Figure 2.11: 2000-2004, in this period the three largest components are presented since they are mostly similar in number of authors and it also helps us to better visualize the evolution of the network.

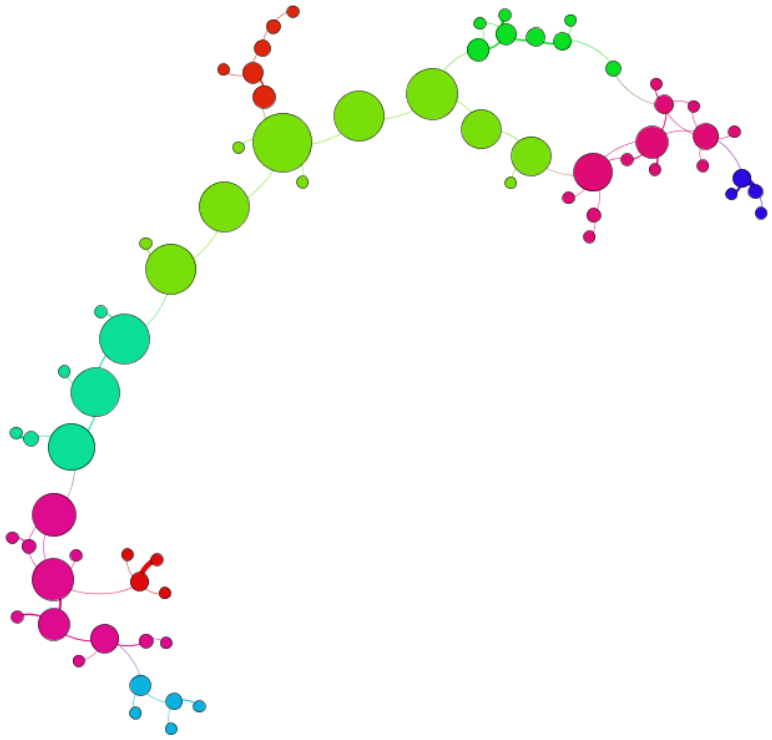


Figure 2.12: 2000-2005; from this time one would notice a shift in the evolution of the network and would observe the presence of the *one* giant component. This giant component consists of 69 authors which accounts for $\sim 35\%$ of the whole network.

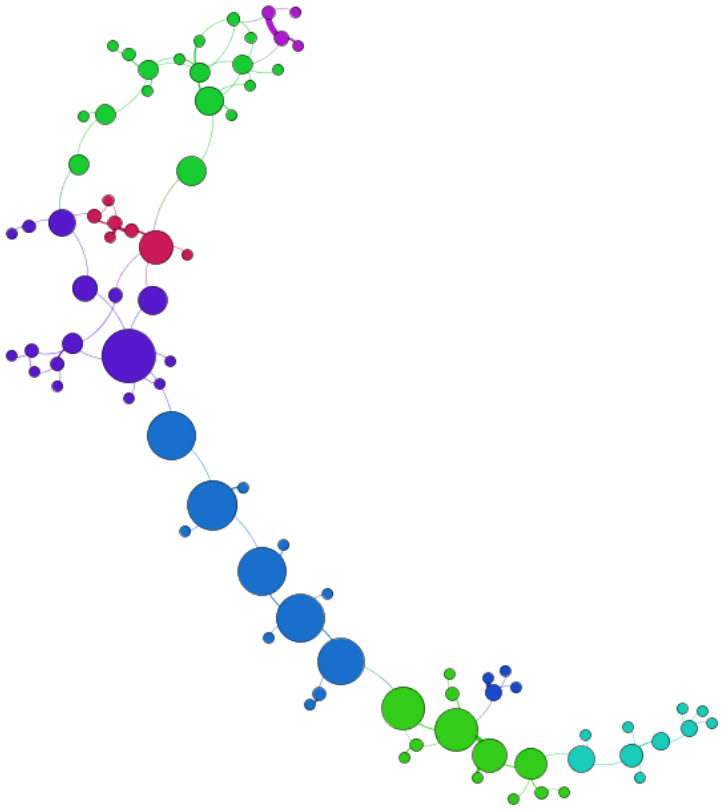


Figure 2.13: 2000-2006; the giant component in this period consists of 83 authors that accounts for $\sim 40\%$ of the whole network.

Figure 2.14: 2000-2007; the giant component in this period consists of 105 authors that accounts for $\sim 47\%$ of the whole network.

Figure 2.15: 2000-2008; the giant component in this period consists of 130 authors that accounts for $\sim 57\%$ of the whole network.

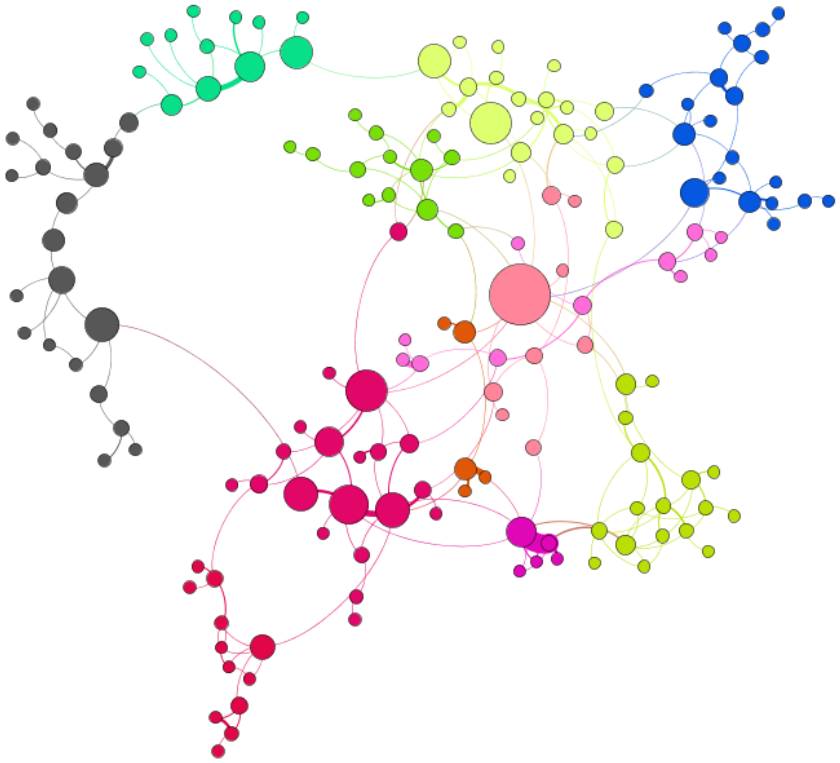


Figure 2.16: 2000-2009; the giant component in this period consists of 164 authors that accounts for $\sim 66\%$ of the whole network.

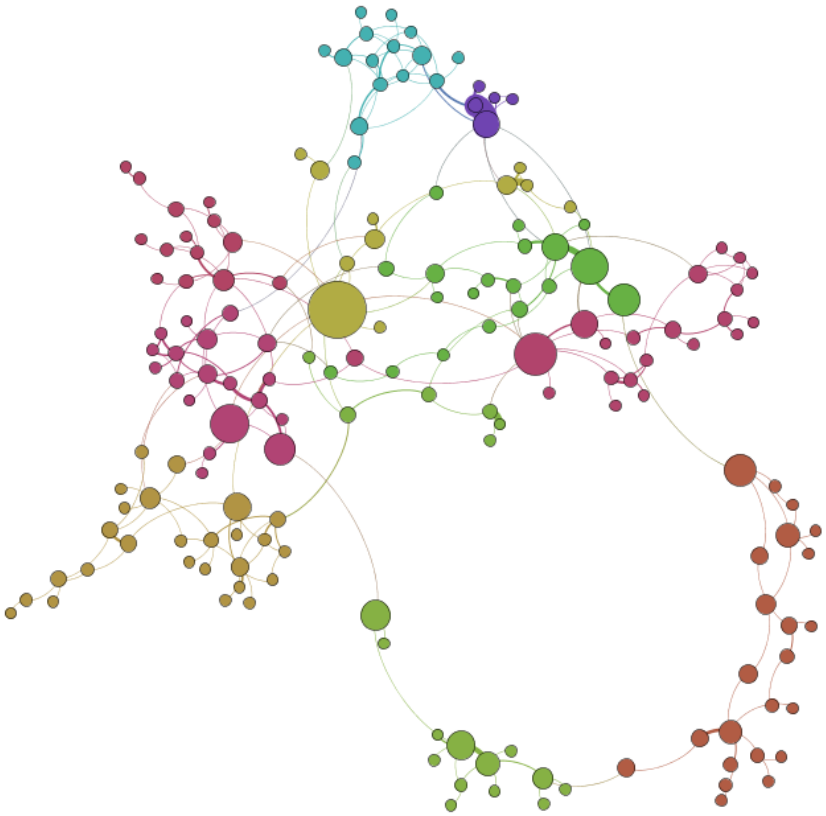


Figure 2.17: 2000-2010; the giant component in this period consists of 172 authors that accounts for $\sim 67\%$ of the whole network.

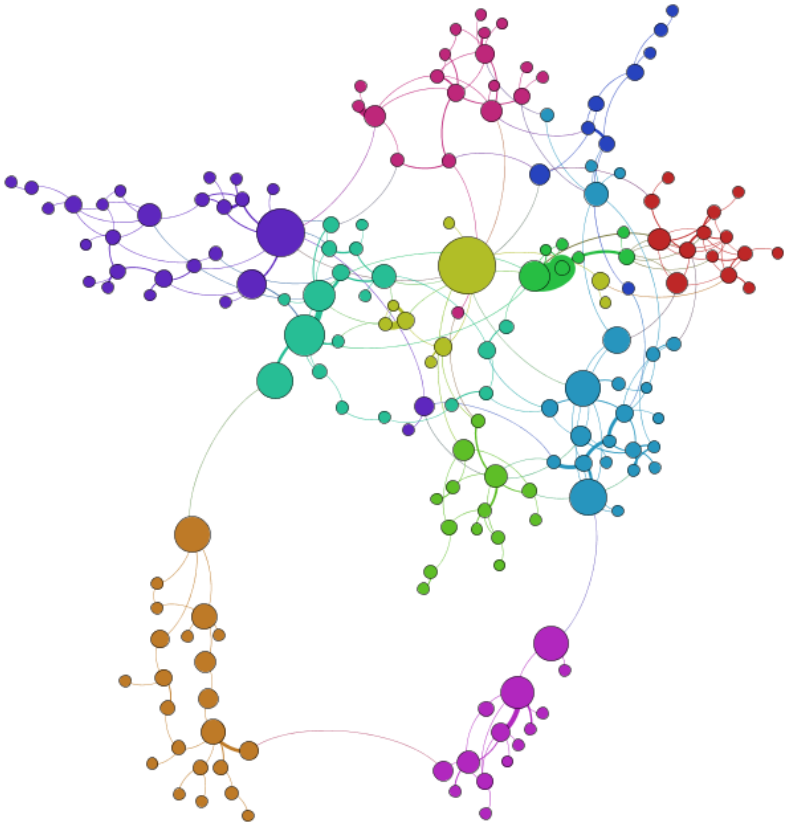


Figure 2.18: 2000-2011; the giant component in this period consists of 178 authors that accounts for $\sim 68\%$ of the whole network.

2.5 The Importance of the Central Nodes

The nodes with a high value of betweenness centrality or low value of closeness centrality are usually called the central nodes. The high value of betweenness for a professor in collaboration networks means that the node/professor has an important role for the diffusion of the information since a lot of shortest paths between pairs of other professors pass through it. The low value of closeness means that the node has the lowest distances to other professors and it is the closest to everyone. Just in order to put an accent on the importance of such nodes in our studied network, some of the most central nodes in the giant component have been removed and some of the metrics have been recalculated in order to see how the absence of such nodes would affect the network.

In HEC Montréal publication network the highest value of betweenness centrality is equal to 4790.8 that belongs to a professor from the department of Management Sciences. We first removed this author and calculated the metrics for the resulting network. The resulting network consists of two components with sizes of 176 and 1 nodes. The average shortest path length is increased from 6.08 to 6.47 which was expected since this author has the biggest value of betweenness and its presence assures the shortest path between lots of pairs of nodes. Besides from this metric there are no other significant changes in other metrics.

The lowest value of closeness centrality is equal to 3.95 that belongs to the very same professor coming from the department of Management Sciences with the highest value of betweenness. As a result, this author is considered the most central and influential one in the network, he does not depend on others for relaying messages [8] since he has an almost direct access to the most parts of the network with the average of 3.95 steps, and finally the removal of this node will result into a bigger average shortest path length in the network.

Chapter 3

Communities

3.1 Community Structure

Many real networks share some common properties such as the small world effect, fat-tailed degree distribution and clustering [23] (each of these aspects were explained in detail in Section 1.3). Another property that they have in common is the "community structure" [23]. In real networks the distribution of edges is inhomogeneous, i.e. the edges within some groups are highly concentrated whereas in some others they are much less concentrated [21].

A group of nodes with a relatively higher density of internal links versus their external links towards other sets of nodes are called communities or clusters [29]. The nodes within these communities are known to share some common properties which depends on the nature of the network [21], as an example in the WWW network (World Wide Web) these communities may correspond to groups of pages on related topics [20].

Social networks are examples of networks with community structure, this is only imaginable since "people naturally tend to form groups" [21]. In the case of collaboration networks the communities may correspond to the groups of people coming from the same departments or working on the same subjects.

3.2 Community Detection Applications

Now that the presence of the community structure in real networks is known, another question could be raised: Why is it useful to find such communities in networks? Revealing

the communities may help better understand how complex networks are organized and how they function [29]. Once the communities are revealed one could extract the structure of the network and also the features and characteristics of its nodes, since the nodes that shape communities not only share edges but also share some common characteristics [50].

Depending on the nature of the network, there would be various applications of community detection, such as improving sales opportunities by finding people with relatively similar profiles in the network of the relationships among customers and products. For instance, retailers like Amazon.com Inc. recommend the same groups of products whiten each community since they share similar profiles [28].

Another interesting application would be finding the central nodes in each community and also finding nodes playing an important role of information transfer between communities. Once the communities and their boundaries are found, nodes having the most links with other members would be the central ones, and nodes sharing edges with other communities are the ones responsible for the connection and knowledge transfer between communities [21].

In the case of collaboration networks, finding the communities will help understand what the profile and the background of the professors belonging to one community are and would also explain what common properties would bring the professors together in a community.

3.3 Community Detection Algorithm

Since there is no clear definition of communities and clusters, it is imaginable that there would be plenty of community detection methods available in the literature [21]. All of these methods usually assume the presence of natural communities or subgroups in networks and all one need to do is to go and find them [40].

The algorithm that is used in this study is the one used in Gephi (an open-source network analysis software). The method is called the Louvain community detection algorithm and has been presented by Blondel et al. in “Fast unfolding of communities in large networks” [10]. This heuristic method is known to “outperform” other known methods “in terms of computational time” and also reveals good quality communities [10].

The objective of this method is to maximize the objective function (modularity (Q)) which is the following equation presented by Newman [37] for the weighted networks:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where A_{ij} is the weight between nodes i and j, k_i and k_j are the weighted degrees of the nodes i and j, c_i and c_j are the communities where nodes i and j are assigned to, m is the total number of the links and $\delta(c_i, c_j)$ is a function that returns 1 whenever c_i and c_j are equal and 0 otherwise [10].

This method works with the weighted network where the edges could have weights, in other words in our studied network if two professors publish together multiple times, it would be taken into consideration in the community detection. The method has two major steps that are repeated until the modularity is maximized, these steps are explained in details:

The first step of the algorithm consists of assigning each node to a different community so that the number of the communities would be equal to the number of the nodes, then the node i is moved to the communities of its neighbours and the gain of the modularity is calculated; if the swap leads to a positive gain, the node will join the community of the node that leads to the highest gain and if there is no positive gain, the node will stay in its original community [10].

In the second step, a new network is created using the communities already detected in the first step. In this new network each community transforms into one new node and the weights between the nodes in the new network is "the sum of the weights of the links among nodes in the corresponding two communities" [10]. Once this networks is completed, the heuristic repeats the whole procedure until there is no further improvements in the modularity [10].

3.4 Homogeneity in Communities

The communities in the studied network are found with the Louvain community detection algorithm [10] as explained in Section 3.3. The results show that at the end of the 12-year

period there are 12 communities in the largest component (Section 2.2.1), which consists of 185 nodes representing about 68% of the whole network.

Each professor present in the network belongs to a department, as a result in a community there could be professors coming from various departments (Section 2.2.3). In order to study the homogeneity of each community the percentage of the members coming from each department is calculated. The largest percentage of professors coming from the same department within each community is presented in Figure 3.1. The x axis represents the community size and the y axis represents the ratio of the largest number of professors coming from the same department divided by the size of that community.

Observing Figure 3.1, it seems that the homogeneity of communities tend to decrease as their sizes grow, meaning that in bigger communities we observe an implication of professors coming from different departments. However, the variation is big which prevents us from a general conclusion.

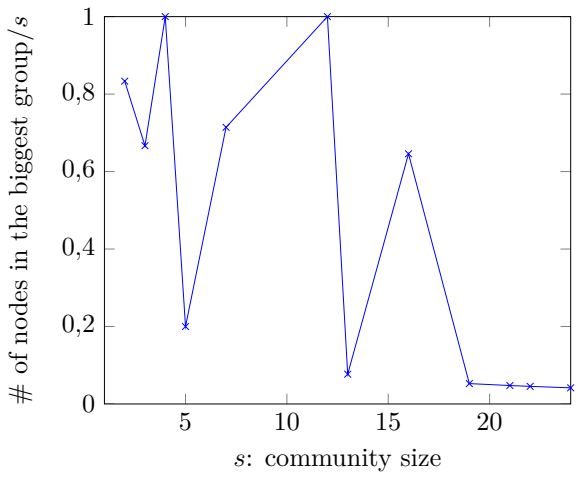


Figure 3.1: Homogeneity of Communities

3.5 Distribution of Departments in Communities

In order to get a more detailed view of the situation of communities, The departments distributions in each of these communities are presented in Figures 3.2 thru 3.13. it should be noted that the analysis is focused only on the giant component at the end of the studied period in 2011 as already illustrated in Figure 2.18.

Some information that are extracted from the following Figures are the frequency of collaborations among departments and also the departments that have never collaborated together which are presented below:

The two departments that have worked the most with others in the same communities are the applied economics and the finance departments.

The member of the applied economics department have frequently published with the member of finance and management sciences departments. On the other hand, the member of the international business department have only been present in the same community with only finance, human resource management and accounting studies departments and they have not collaborated with other departments.

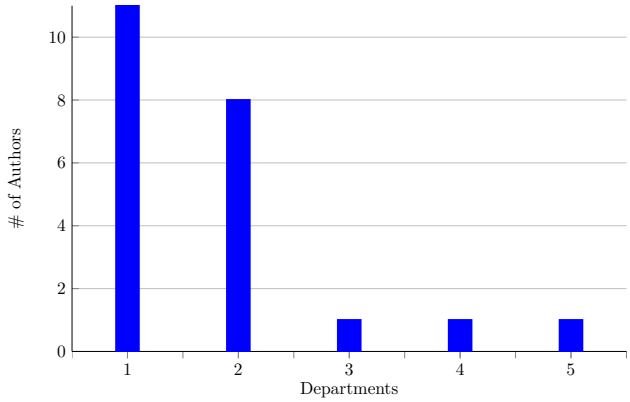


Figure 3.2: Community no.1 (1:Logistics and Operations Management 2:Management 3:Marketing 4:Applied Economics 5:Finance)

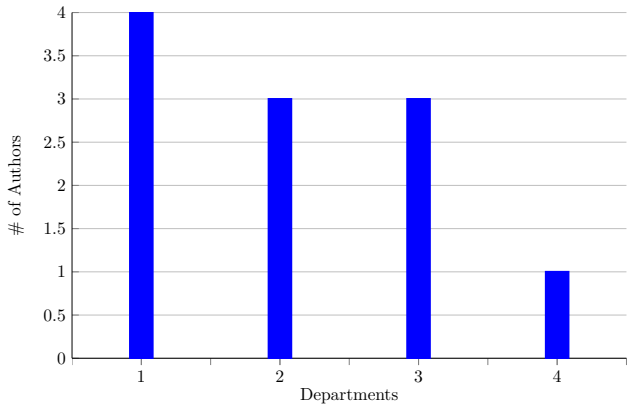


Figure 3.3: Community no.2 (1:Applied Economics 2:Finance 3:Management Sciences 4:Marketing)

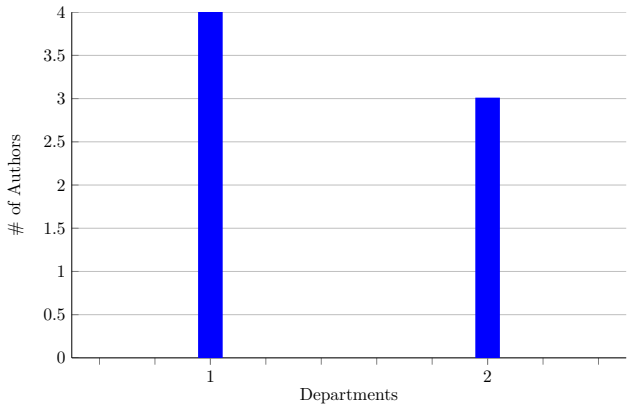


Figure 3.4: Community no.3 (1:Logistics and Operations Management 2:Management Sciences)

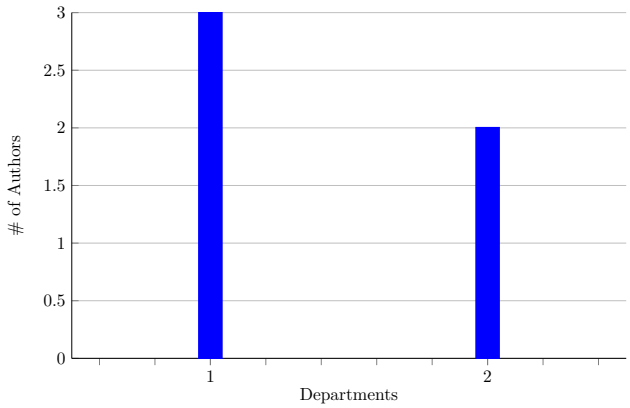


Figure 3.5: Community no.4 (1:Management Sciences 2:Marketing)

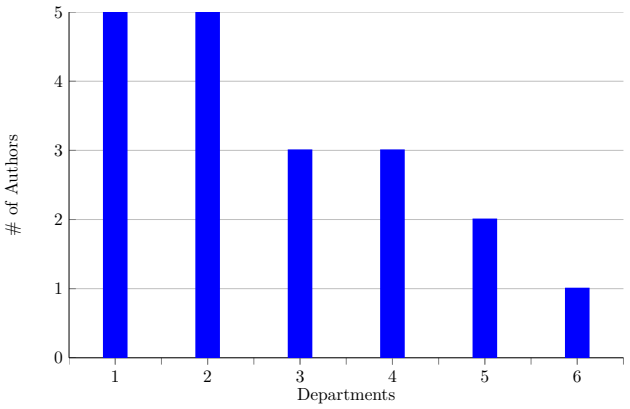


Figure 3.6: Community no.5 (1:Applied Economics 2:Management Sciences 3:Finance 4:Information Technologies 5:Marketing 6:Accounting Studies)

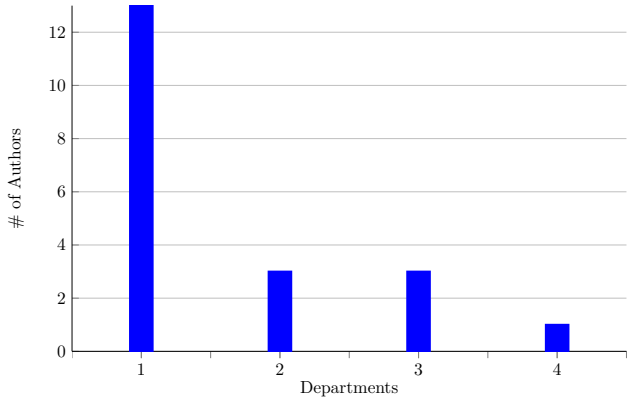


Figure 3.7: Community no.6 (1:Human Resource Management 2:Management Sciences 3:Accounting Studies 4:Management)

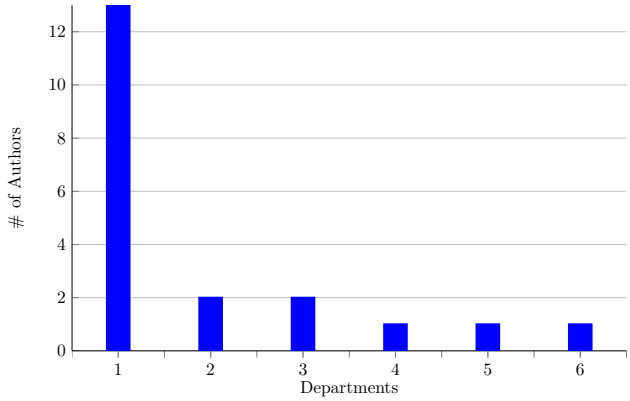


Figure 3.8: Community no.7 (1:Information Technologies 2:Management 3:Organizational Management and Leadership 4:Human Resource Management 5:Applied Economics 6:Finance)

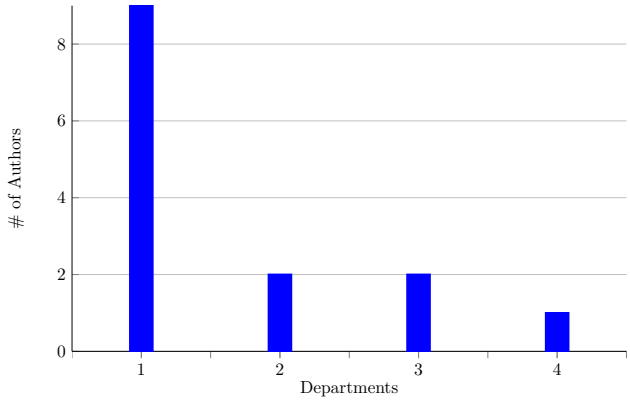


Figure 3.9: Community no.8 (1:Logistics and Operations Management 2:Accounting Studies 3:Marketing 4:Applied Economics)

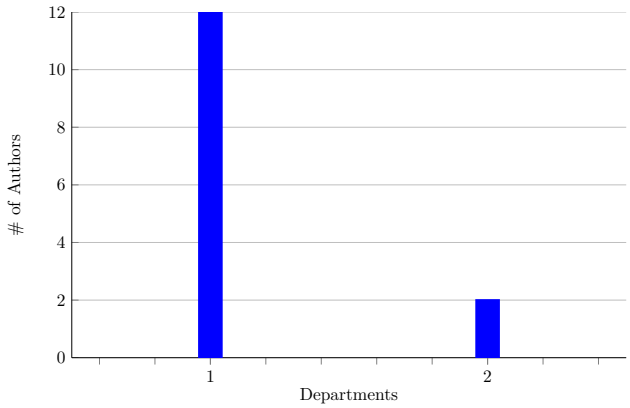


Figure 3.10: Community no.9 (1:Marketing 2:Applied Economics)

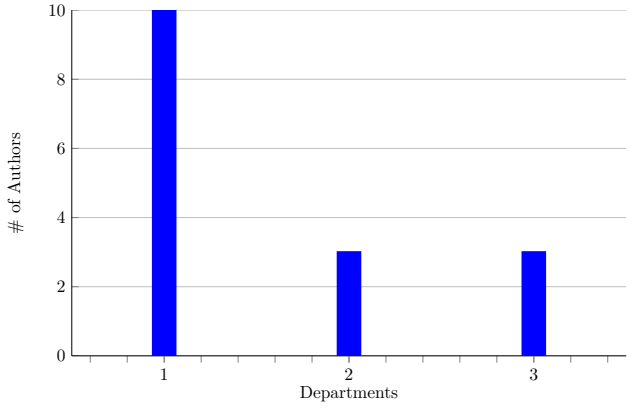


Figure 3.11: Community no.10 (1:Finance 2:Management Sciences 3:Applied Economics)

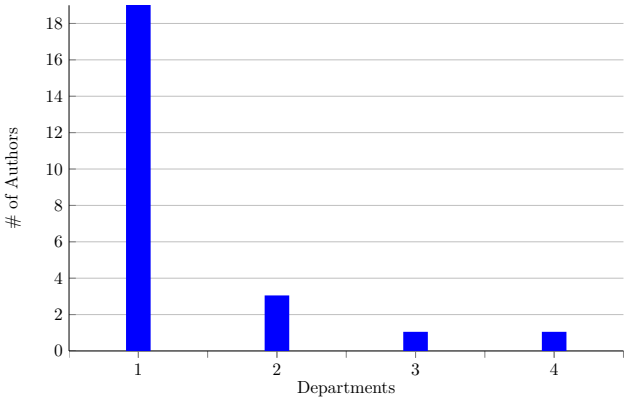


Figure 3.12: Community no.11 (1:Accounting Studies 2:Logistics and Operations Management 3:Applied Economics 4:Finance)

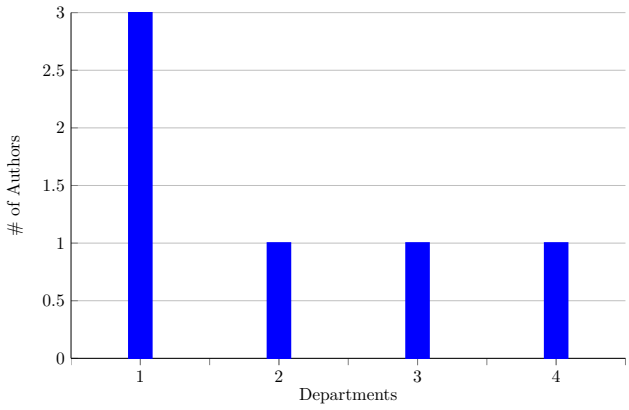


Figure 3.13: Community no.12 (1:Finance 2:Accounting Studies 3:Human Resource Management 4:International Business)

3.6 Collaborations Between Communities

After having studied the characteristics of each community, it would be interesting to also study the collaboration amongst them. This way one would observe which communities are collaborating more often together and would also find the communities with the biggest value of betweenness centrality and smallest value of closeness centrality which means they would act as the most influential and central communities in the network.

Betweenness centrality is the frequency of the number of times a node "falls between pairs of other nodes on the shortest or geodesic paths connecting them" [22]. The nodes with a high degree of betweenness centrality are "strategically located" on the communication paths between other nodes [7], have the potential of coordinating the "group processes" [14] and are responsible for the "maintenance of communication" [45].

Closeness centrality is the value that measures the sum of distances between a node to all other nodes [22]. The communication with other nodes in the network would cost the least and would also take the least time for the most central nodes [44], therefore it would be interesting to find such nodes in the network.

In order to find the central communities explained above, we have constructed a new network where the new nodes are the communities in the giant component found in Section 3.4, and the new edges are the weighted links between these communities. This network is presented in Figure 3.14 where the node sizes are proportional to betweenness centrality (the bigger nodes have higher betweenness values), the nodes are coloured according to closeness centrality (the darker coloured nodes have bigger closeness value) and finally the edges are proportional to the links weights (the thicker edges have the bigger weights).

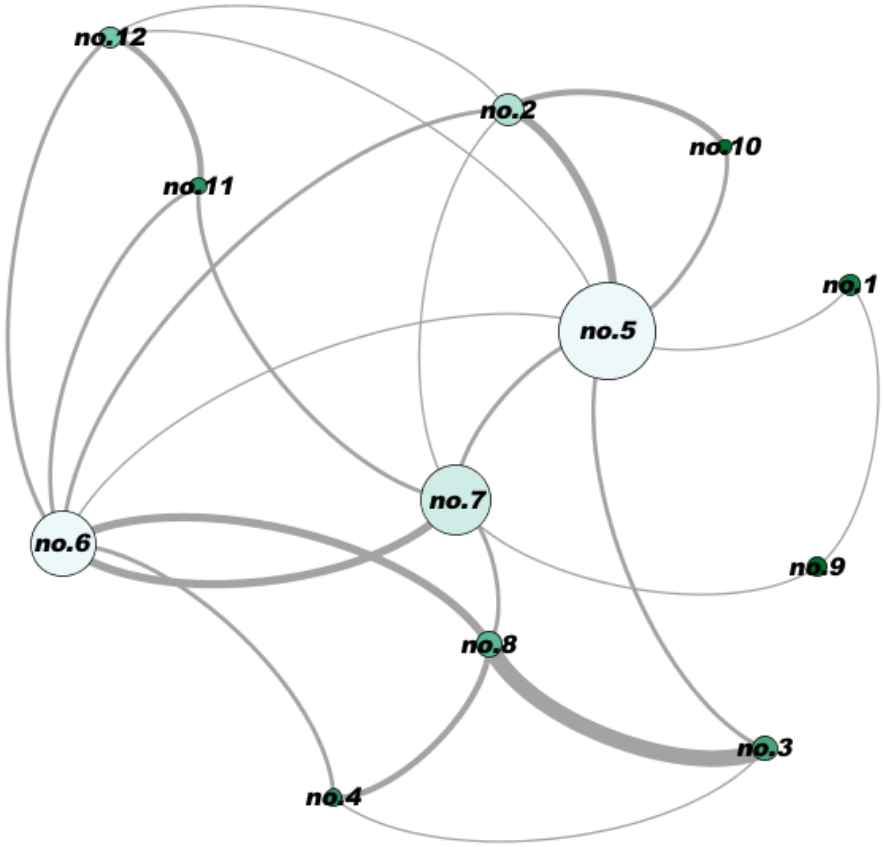


Figure 3.14: Collaborations Among Communities

As seen in Figure 3.14, the two communities that have the most collaborations together are communities no.3 (Figure 3.4) and no.8 (Figure 3.9) with the weighted link of 8. Community no.5 (Figure 3.6) has the highest value of betweenness centrality meaning that this community has the most important role for the diffusion of the information in the network since a lot of shortest paths between pairs of communities pass through this one. Community no.5 and community no.6 (Figure 3.7) have the lowest value of closeness centrality (1.364), meaning that these two have the lowest distances to all other communities and that they are the closest to everyone.

Also, community no.5 holds an important position since it has the biggest value of betweenness centrality and the lowest value of closeness centrality simultaneously. This community consists of 19 professors mostly from six different quantitative departments and is the most central and influential one in the studied network. This central community does

not depend on others for relaying messages [8], in other words it has an almost direct access to the most parts of the network with an average of 1.36 steps.

Chapter 4

Preferential Attachment

The degree distribution of our studied network is fat-tailed as discussed in Section 2.2.5. This aspect differentiates the network from random networks where nodes are connected together randomly regardless of their degrees [18, 49]. With the "increasing evidence" [26] that real networks do not behave as random ones, new models have been presented to better understand them. Models which study the networks dynamically instead of statically allows us to study the evolution of such networks over time [26]. These evolving networks are found to have two characteristics: "Preferential attachment and growth" [4, 26]. The first characteristic states the fact that there are rules that control the linking between nodes, i.e. preferential attachment which is the assumption that in scale-free networks the higher degree nodes acquire new nodes faster than lower degree ones [5, 16, 26]. The second characteristic points out that these networks expand over time with the addition of new nodes and edges [5, 26].

To investigate more on how the preferential attachment is affecting the evolution of our network, in the following sections we study the behaviour of the new nodes entering the network, the effect of the previous collaborations on repeated collaborations, the effects of departments and number of common neighbours on co-authorship and finally how preferential attachment differently acts at the community level versus at the node level.

4.1 Behaviour of New Nodes

First we inspect how new nodes link to the already existing nodes when they first appear in the network. A node is considered a new one if it has not been present in the network in the previous years, and an existing node is a professor that has published at least once in the previous years. In order to do so, among the 12 years record of publications in the database, we would use the first 11 years to construct the network and will then analyze the behaviour of the network in the last remaining year. Yet, before starting the analysis we would make some assumptions as suggested by Newman [33]. In a similar work he assumes that an active professor should have already published at least once in the first initial years and any nodes or edges appearing in the last year would be new. As a result, at the end of the period we would have a complete network. Another assumption is that the exceptions such as "established professors" [33] who have not published in the initial years are assumed not to be significantly large to bias our study. Finally, there are some professors leaving the network (for various reasons such as retirement etc.) and it is assumed that it is not necessary to remove them from the dataset as the error caused by this would not significantly affect the study of preferential attachment. Newman argues that the correlation we are going to study would only be "weakened by this error and not strengthened" [33].

After having constructed the 11-year network, we will investigate the behaviour of new nodes entering the database during only the last year. To do so we calculate the probability $\Pi(k)$ that a new professor publishes an article with an old professor with degree k, and then verify whether $\Pi(k)$ is dependant or independent of k. In the case of preferential attachment one would expect that this probability would be dependant on k, i.e. the nodes would not connect randomly and would also be an evidence that the degree distribution follows a power-law distribution [15, 27].

In order to find the probability $\Pi(k)$, Barabàsi et al. [6] calculate the difference of the degrees of the professors between the end of the year 11 and year 12, called Δk. They would then plot Δk as a function of k. The resulting plot shows the probability $\Pi(k)$ which is presented in Figure 4.1. This probability is expected to "grow as k^{ν}" with $\nu > 0$ [47]. In order to better visualize the presence of preferential attachment, the integral of $\Pi(k)$ which is "the cumulative preferential attachment, $K(k)$" is plotted in Figure 4.2.

In the study conducted by Barabàsi et al. [6], the $K(k)$ function is non-linear. As a result, they conclude that the probability $\Pi(k)$ is dependant on k and thus $\nu \neq 0$ [6, 47]. In our study as seen in Figure 4.1, the probability of a professor acquiring new collaborators increases almost linearly as k increases, meaning that professors having more previous collaborators, have a higher chances of acquiring new collaborators. This probability suddenly decreases for large k. Newman [33] argues that this behaviour is normal because "no one can collaborate with an infinite number of people in a finite period of time" [33].

Comparing Figure 4.1 with the degree distribution previously shown in Figure 2.2, one would observe that in both cases the functions start to decrease at almost the same degree of collaboration, i.e. the preferential attachment starts to decrease at $k = 5$ whereas the degree distribution starts to decrease at $k = 4$. This minor difference could be attributed to the sub-linear behaviour of the preferential attachment which "gives rise to a stretched exponential cutoff in the resulting degree distribution" [27]. The fact that both functions start to fall at almost the same degree of collaboration is similar to what Newman stated in a comparable study that this phenomenon is "a support to the theory that preferential attachment is the origin of power-law" [39].

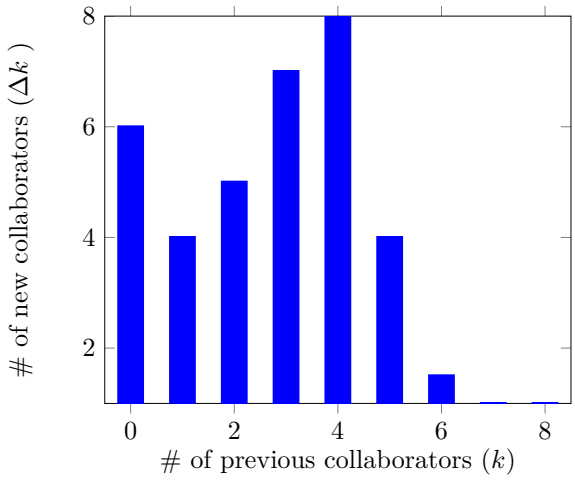


Figure 4.1: The Probability that an Author with Degree k Acquires a New Collaborator

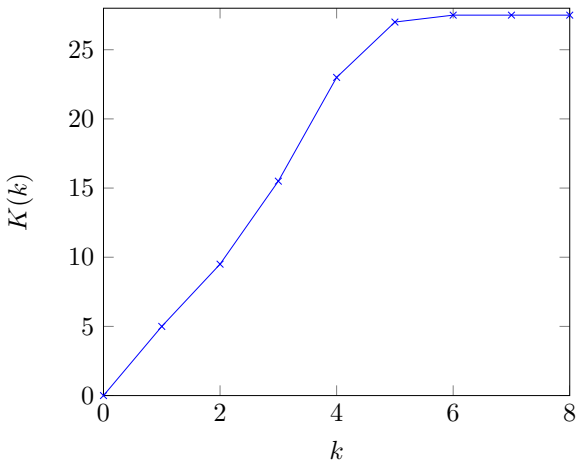


Figure 4.2: Cumulated Preferential Attachment for New Professors Entering the Network in the Last Year

4.2 Repeat Collaborations

Another interesting aspect to study in the network is whether the number of previous collaborations between two scientists during the first 11 years would raise the probability of co-authoring another article in the last year. In Figure 4.3 the relative probability of two professors co-authoring with m previous collaborations is presented. According to this plot there is no evidence of the influence of the number of previous collaborations on the decision of co-authoring again in the last year. The result is somehow surprising and in order to investigate more on this subject, the frequency of the number of repeated collaborations among pairs of professors in the whole 12-year network is presented in Figure 4.4.

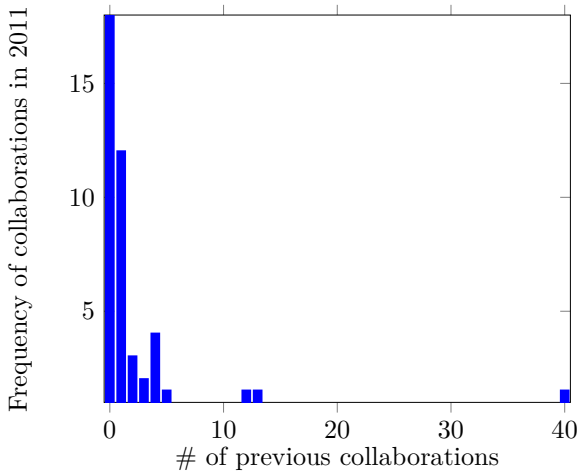


Figure 4.3: The Probability that Two Authors with m Previous Collaborations Co-author Another Article in the Last Year

As we can see in the following figure, in the 12-year network the repeated collaborations are not very common. Setting aside the 79% of the articles that have been published with only one author in these 12 years, 200 pairs of professors (71% of the total repeated collaborations) have published together only once and 42 of them (14% of the total repeated collaborations) have published together only twice.

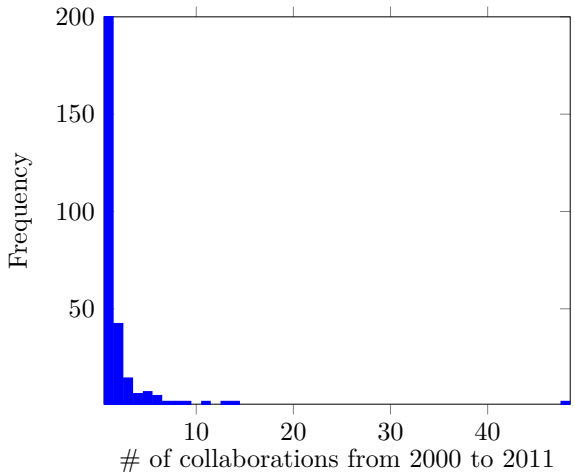


Figure 4.4: The Distribution of the Number of Repeat Collaborations Among Pairs of Nodes in the 12-year Network

So far it has been observed that the number of previous collaborations between authors do not seem to affect the decision of co-authorship in our network. Another interesting

question is whether this decision is affected by the departments the professors belonged to, or by the number of neighbours they have in common. In the upcoming Sections 4.3 and 4.4 these two aspects are studied.

4.3 Effects of Departments on Preferential Attachement

In Table 2.III the complete statistics for different departments of HEC Montréal have been presented. There are 12 different departments and the department each professor belongs to is also known. With this given information one needs to verify whether the publications have been done between the professors from the same department or not. We have verified this aspect at two levels: first for the complete database until the end of year 2011 and the second one for only the last year 2011. The result has been presented in Figure 4.5.

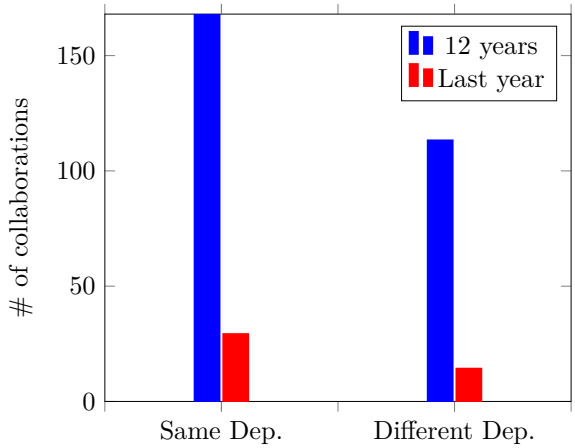


Figure 4.5: The Probability that Two Authors from the Same Department Co-author

As seen in Figure 4.5, for both time periods, professors tend to collaborate more within the department they belong to. For the complete network at the end of the year 2011 the percentage of publications within the same departments is 60% versus 40% between departments. Same percentage for the network of only year 2011 is 67% versus 33%. This is another evidence that there are aspects governing the decision of co-authership amongst the professors.

4.4 Effects of Number of Common Neighbours on Preferential Attachment

As previously discussed in Section 1.2, one of the characteristics of social networks which differentiates them from non-social networks is the presence of clustering coefficient which is the likelihood that two nodes that have a node in common connect with each other. The value of the clustering coefficient for each year and for the complete network is presented in Table 2.I . Now that is known that two authors are more likely to co-author when they have another node in common, an interesting question would be: Does this likelihood increase as the number of common nodes (common neighbours) increase? To answer this question Newman [33] introduces the probability R_m which is the relative probability that two scientists with m common neighbours in the previous years, collaborate in the upcoming years. He presents the following equation to calculate R_m:

$$R_m = \frac{1}{n_m(t)} P_m(t) \frac{1}{2} N(t)[N(t) - 1]$$

where $n_m(t)$ is "the number of pairs with m mutual neighbour in previous years", $N(t)$ is "the current number of authors in the network" and finally $P_m(t)$ is "the probability that two scientists connected by a link added at time t have m mutual neighbours" [33]. The relative probability of collaboration between professors (R_m) in year 2011 is presented for different numbers of previous common neighbours, m, in Figure 4.6:

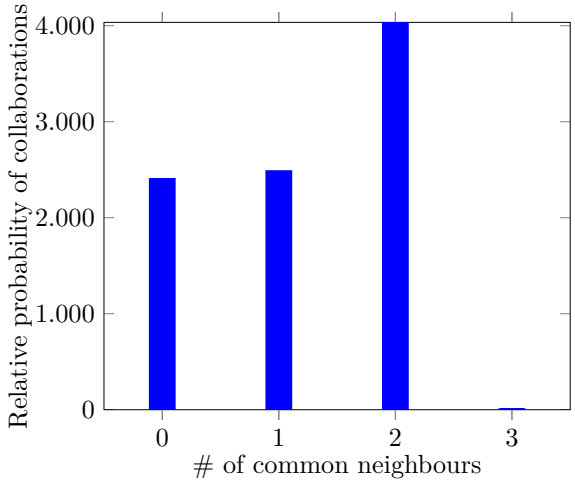


Figure 4.6: The Probability that Two Authors with m Common Neighbours Co-author

It is observed that the professors with two common neighbours are much more likely to collaborate compared to the professors having no or only one common neighbours. An interesting point concerning Figure 4.6 is that the relative probability of collaboration for a pair of scientists having three common neighbour is zero. An underlying reason would be the fact that we are working with a relatively small portion of the network, i.e. among the 43 edges (there would be an edge between two professors if they have co-authored an article together) in 2011, only 25 of them are new links (a brand new collaboration between pairs of scientists that has not been present in the previous years) and thus included for the preferential attachment calculations. Newman [33] also raises an interesting point in a similar work that the number of pairs of scientists having this much common neighbours that have not already collaborated after 11 years is really small and less likely to lead to new edges.

4.5 Preferential Attachment at the Community Level

Until now we have studied the presence of preferential attachment at the nodes level (Sections: 4.1, 4.2, 4.3 and 4.4). An interesting question would be whether the same process is governing the evolution of the network at the community level. Pollner et al. [43] study the process of preferential attachment at a higher level than nodes, i.e. communities. In Figure 4.7 the distribution of the size of the communities for the cumulative network at the end of year 2011 is presented. As shown in the figure, the size distribution is fat-tailed, i.e. there is a large number of small communities compared to a smaller number of larger communities.

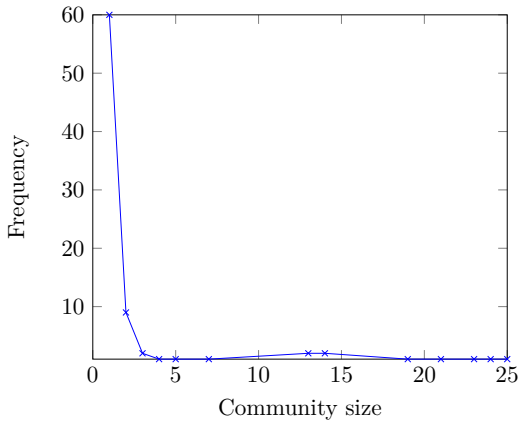


Figure 4.7: Community Size Distribution

Knowing that the community size distribution is fat-tailed, one could raise the question what would be the size of the target communities that the nodes not belonging to any community would like to attach to. To investigate more on this subject, we constructed the network for the first 11 years and then observed how the nodes reacted in the remaining last year in 2011. As a result, we first identified the nodes belonging to none of the communities in the initial 11 years and then verified whether in the last year they connected to a community or not, and if the answer is yes what was the size of that target community in the previous year?

The challenge is how to identify the picture of the actual community in 2011, in the last time stamp since there are several possible scenarios that could happen for the communities in evolving networks. Palla et al. [42] name these different possibilities: "Growth, death, birth, merging, contraction and splitting". Also, Fortunato [21] argues that the detection of the communities statically is still "controversial", so the challenge with community detection in dynamic networks is understandable. Palla et al. [42] suggest a method for identifying the communities at different time stamps. Their method detects the community $C(t+1)$ at time $t+1$ and then finds the community that has the most overlap with this community at time t; the resulting community would be the image of $C(t+1)$ at time t. Using the same approach in section 4.4, we calculated the relative probability that a node not belonging to any community in the first 11 years, choses a community with m members. This relative probability is presented in the following Figure 4.8:

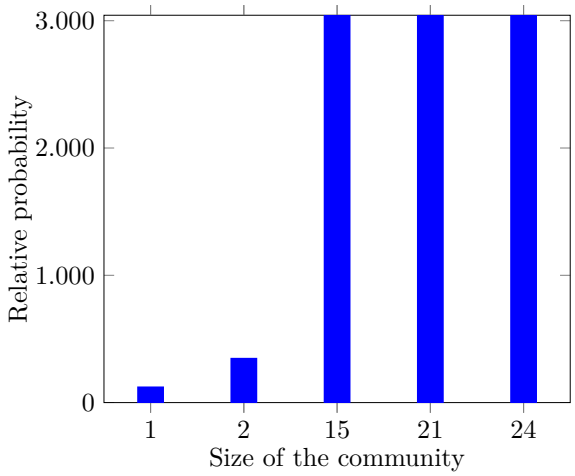


Figure 4.8: The Relative Probability that a Node Belonging to no Communities Joins a Community in the Upcoming Year

As shown in the figure, the bigger communities are more likely to attract new professors. This phenomenon explains the fat-tailed distribution of community sizes presented in Figure 4.7 and also confirms the presence of preferential attachment at the communities level.

General Conclusion

Considering the fact that the collaboration networks are getting more attention recently due to their capability of affecting scientific practices [32] and also the lack of dynamical approaches towards such networks in the literature, we have been motivated to conduct this research. The collaboration network of the HEC Montréal's professors is used for this work which is the first time this database is used to study the publications patterns at the school.

With the help of the network science, we established a graph network presentation of the publication collaboration and studied its evolution over time using descriptive statistics, social network analysis and dynamic network analysis methods.

Comparing to the results of the similar works discussed in the literature, the network of HEC Montréal's publications shares some common properties with other collaboration networks: 1) It forms a small-world where each pair of the nodes are connected to each other through "a short chain of acquaintances" [49], 2) The clustering coefficient is present and thus, two authors having a common neighbour are more likely to co-author, 3) The degree distribution is fat-tailed which confirms the presence of preferential attachment, and finally 4) The presence of community structure i.e., presence of the group of nodes with a relatively higher density of internal links versus their external links towards other sets of nodes are called communities or clusters [29].

By studying the network cumulatively and dynamically, it is observed that the average degree in our studied network is increasing over time, on the other hand the average path length, average clustering coefficient and the relative size of the largest component only increase at the beginning of the studied period and tend to have a more stable value in the last years where the network has been shaped.

After uncovering the community structure in the studied network, the distribution of departments in each of these communities and therefore their homogeneities were studied. The most central communities were presented and also the communities with frequent collaborations together were detected.

With the study of the preferential attachment, it is known that, new nodes entering the network for the first time are more likely to attach to the nodes having a bigger degree and are also more likely to attach to bigger communities. The results also show that a pair of nodes having three common neighbours are about twice more likely to collaborate compared to a pair of nodes with zero common neighbours. Finally, it is known that there is more tendency among professors in the same department to collaborate compared to the ones coming from different departments.

As a future work we propose constructing the bipartite graphs [30] which is a graph where its nodes are divided into two sets of nodes and there are no links among the nodes in each set. In our collaboration network one set would be the papers and the other would be the authors. Each paper is linked to the authors that have published it and each author is linked to the papers that has published. Including such presentation of the network would increase the calculation time and also "would increase the fidelity of the model" [6]. We also recommend examining exact methods to detect communities as future research.

Based on the extensive literature review, it is known that the dynamical approaches towards networks are still in their infancy. As Carley states: "To create a truly dynamic network theory we need to create the equivalent of a quantum dynamics for the socio-cognitive world, where the fundamental entities, the people, unlike atoms, have the ability to learn" [13]. The contribution of this work is that it combines various fragmented social network analysis techniques available in the literature to analyze the academic collaboration network.

Bibliography

[1] Alexanderson G. About the cover: Euler and Königsberg's bridges: A historical view. Bulletin of the american mathematical society 2006;43(4):567–573.

[2] Barabàsi A. Linked: How everything is connected to everything else and what it means for business, science, and everyday life. Plume Editors 2002;.

[3] Barabàsi A. Network Science, 2012.

[4] Barabási A, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–512.

[5] Barabàsi A, Albert R, Jeong H. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: Statistical Mechanics and its Applications 2000;281(1):69–77.

[6] Barabàsi A, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 2002;311(3):590–614.

[7] Bavelas A. A mathematical model for group structures. Human organization 1948;7(3):16–30.

[8] Bavelas A. Communication patterns in task-oriented groups. Journal of the acoustical society of America 1950;.

[9] Biggs N, Lloyd E, Wilson R. Graph Theory 1736-1936. Oxford University Press, 1976.

[10] Blondel V, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008;2008(10):P10008.

[11] Börner K, Sanyal S, Vespignani A. Network science. Annual review of information science and technology 2007;41(1):537–607.

[12] Brandes U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 2001;25(2):163–177.

[13] Carley K. Dynamic network analysis. In: Dynamic social network modeling and analysis: Workshop summary and papers. Citeseer; 2003. p. 133–145.

[14] Cohn B, Marriott M. Networks and centres of integration in indian civilization. Journal of Social Research 1958;1(1):1–9.

[15] Dorogovtsev S, Mendes J, Samukhin A. Structure of growing networks with preferential linking. Physical Review Letters 2000;85(21):4633.

[16] Dorogovtsev SN, Mendes JF. Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press, 2013.

[17] Erdős P, Rényi A. On random graphs. Publicationes Mathematicae Debrecen 1959;6:290–297.

[18] Erdős P, Rényi A. On the strength of connectedness of a random graph. Acta Mathematica Hungarica 1961;12(1):261–267.

[19] Euler L. Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae 1741;8:128–140.

[20] Flake G, Lawrence S, Giles C, Coetzee F. Self-organization and identification of web communities. Computer 2002;35(3):66–70.

[21] Fortunato S. Community detection in graphs. Physics Reports 2010;486(3):75–174.

[22] Freeman L. Centrality in social networks conceptual clarification. Social networks 1979;1(3):215–239.

[23] Girvan M, Newman M. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 2002;99(12):7821–7826.

[24] Granovetter M. The strength of weak ties. American journal of sociology 1973;:1360–1380.

[25] Grossman J. The evolution of the mathematical research collaboration graph. Congressus Numerantium 2002;:201–212.

[26] Jeong H, Néda Z, Barabàsi A. Measuring preferential attachment in evolving networks. EPL (Europhysics Letters) 2003;61(4):567.

[27] Krapivsky P, Redner S, Leyvraz F. Connectivity of growing random networks. Physical review letters 2000;85(21):4629.

[28] Krishnamurthy B, Wang J. On network-aware clustering of web clients. In: ACM SIGCOMM Computer Communication Review. ACM; volume 30; 2000. p. 97–110.

[29] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Physical review E 2009;80(5):056117.

[30] M.E.J. Newman SS, Watts D. Random graphs with arbitrary degree distributions and their applications. Physical Review E 2001;64(2):026118.

[31] Milojević S. Modes of collaboration in modern science: Beyond power laws and preferential attachment. Journal of the American Society for Information Science and Technology 2010;61(7):1410–1423.

[32] Moody J. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. American sociological review 2004;69(2):213–238.

[33] Newman M. Clustering and preferential attachment in growing networks. Physical Review E 2001;64(2):025102.

[34] Newman M. Scientific collaboration networks. i. network construction and fundamental results. Physical review E 2001;64(1):016131.

[35] Newman M. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical review E 2001;64(1):016132.

[36] Newman M. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 2001;98(2):404–409.

[37] Newman M. Analysis of weighted networks. Physical Review E 2004;70(5):056131.

[38] Newman M. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National academy of Sciences of the United States of America 2004;101(Suppl 1):5200–5205.

[39] Newman M. Who is the best connected scientist? a study of scientific coauthorship networks. In: Complex networks. Springer; 2004. p. 337–370.

[40] Newman M. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 2006;103(23):8577–8582.

[41] Newman M, j. Park . Why social networks are different from other types of networks. Physical Review E 2003;68(3):036122.

[42] Palla G, Barabàsi A, Vicsek T. Quantifying social group evolution. Nature 2007;446(7136):664–667.

[43] Pollner P, Palla G, Vicsek T. Preferential attachment of communities: The same principle, but a higher level. EPL (Europhysics Letters) 2006;73(3):478.

[44] Sabidussi G. The centrality index of a graph. Psychometrika 1966;31(4):581–603.

[45] Shimbel A. Structural parameters of communication networks. The bulletin of mathematical biophysics 1953;15(4):501–507.

[46] Soffer S, Vázquez A. Network clustering coefficient without degree-correlation biases. Physical Review E 2005;71(5):057101.

[47] Sowe SK, Stamelos IG, I. Samoladas IM. Emerging Free and Open Source Software Practices. IGI Global, 2008.

[48] Valderrama-Zurián JC, González-Alcaide G, Valderrama-Zurián FJ, Aleixandre-Benavent R, Miguel-Dasit A. Coauthorship networks and institutional collaboration in revista española de cardiología publications. Revista Española de Cardiología (English Edition) 2007;60(2):117–130.

[49] Watts D. Small worlds: the dynamics of networks between order and randomness. Princeton university press, 1999.

[50] Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes. arXiv preprint arXiv:14017267 2014;.