A bibliometric analysis of Quebec’s PhD students’ contribution to the advancement of knowledge.

Vincent Larivière

School of Information Studies, McGill University
Montreal, January 2010

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

© Vincent Larivière 2010 All rights reserved
À mes grands-pères.
Abstract

Graduate students are an important part of the academic workforce. However, little is known on their overall contribution to science. Using the participation in Web of Science indexed peer-reviewed publications of the complete population of doctoral students in Quebec over the 2000-2007 period (N=27,393), this thesis achieves three main contributions to the advancement of knowledge in the fields of information science, sociology of the scientific community and sociology of higher education.

The first contribution is a technical one and involves the creation of an algorithm that allows the automatic attribution of a large proportion of individual researchers’ papers. Indeed, using the patterns found in Quebec university researchers’ use of keywords, cited references and discipline of publication, the algorithm automatically attributes or rejects at least one scientific paper to 88% of doctoral students.

The second contribution is to provide a large-scale analysis of doctoral students’ socialization to research, using the percentage of doctoral students who have published at least one paper during their program as an indicator. It shows that this integration varies greatly among disciplines, with students in the natural and medical sciences being more integrated into research than their colleagues of the social sciences and humanities. Collaboration is an important component of this socialization: disciplines in which student-faculty collaboration are higher are also those in which doctoral students are the most integrated into research. Access to research funds also influences doctoral students participation in peer-reviewed papers, as specialties where professors receive greater research funds are also those where students are the most likely to publish. Although the papers to which doctoral students contribute are most often written in collaboration, they are less likely to be the result of international collaboration. Such socialization to research is also positively linked with students’ degree completion and the likelihood of a subsequent career in research.

Finally, the third contribution of this thesis is to measure the percentage of the research output of the research system produced by doctoral students. It provides evidence that, for all disciplines combined, PhD students account for 33% of the publication output of the province, a percentage that is considerably higher than that of Quebec hospital researchers taken together and more than 5 times higher than that of federal and industrial researchers of the province. In terms of scientific impact, papers to which doctoral students have contributed obtain significantly lower citation rates than other Quebec papers to which they have not contributed, although the average impact factor of the journals in which they publish is significantly higher. This suggests that the scientific impact of doctoral students’ papers may suffer from
a Matthew Effect, the sociological phenomenon observed by which recognition for discoveries is more easily attributed to well known scientists than to others less known.

Overall, this interdisciplinary thesis provides a significant insight into the extent, the context and the effect of socialization to research in the PhD curriculum, as well as a better understanding of the importance of doctoral students’ scientific contributions within Quebec’s research system. These findings should be of great interest to university administrators as well as for research councils and the science policy community in general.

Keywords: bibliometrics; scientometrics; doctorate; PhD; students; higher education; universities; Quebec; Canada.
Sommaire

Les étudiants gradués comptent pour une part importante de la main d'œuvre académique. Toutefois, nous ne savons que très peu de choses sur leur contribution globale à l’avancement des connaissances. À partir des articles publiés dans des revues à comités de pairs — et indexés dans le *Web of Science* — par la population complète des étudiants au doctorat au Québec entre et 2000 et 2007 (N=27,393), cette thèse effectue trois contributions principales à l’avancement des connaissances en sciences de l'information et en sociologie de la communauté scientifique et de l'enseignement supérieur.

La première contribution est de nature technique et consiste en la création d'un algorithme qui permet l'attribution automatique à un chercheur d'un pourcentage important de ses articles scientifiques. En effet, en utilisant les régularités trouvées dans les mots-clés, références citées et la discipline de publication des chercheurs universitaires québécois, cet algorithme permet l’attribution ou le rejet automatique d’au moins un article à 88% des étudiants de doctorat.

La seconde contribution est l’analyse à grande échelle de la socialisation des doctorants à la recherche, en utilisant comme indicateur le pourcentage d’étudiants au doctorat qui ont publié au moins un papier au cours de leur programme. Les données montrent que cette intégration varie considérablement entre les disciplines : les étudiants des sciences naturelles et médicales étant plus intégrés à la recherche que leurs collègues des sciences sociales et humaines. La collaboration est un élément important de cette socialisation: les disciplines dans lesquelles la collaboration doctorant-professeur est la plus élevée étant celles où les doctorants sont les plus intégrés dans la recherche. L’accès à des fonds de recherche influence également la participation des étudiants à des publications; les spécialités où les professeurs reçoivent davantage de fonds étant également celles où les étudiants sont plus susceptibles de publier. Bien que les documents auxquels ont contribué les doctorants soient pratiquement tous écrits en collaboration, ils sont moins souvent le résultat d’une collaboration internationale. Cette socialisation à la recherche est également liée de façon positive avec l’obtention du diplôme et la poursuite d’une carrière en recherche.

Enfin, la troisième contribution de cette thèse est la mesure de l’importance, dans l’ensemble de la recherche québécoise, des résultats de recherche auxquels des étudiants de doctorat ont contribué. On y constate que, toutes disciplines confondues, les doctorants ont participé à 33% de la production scientifique de la province, un pourcentage considérablement plus élevé que celui des chercheurs en milieu hospitalier de la province combinés et plus de 5 fois supérieur à celui des
chercheurs du gouvernement fédéral et du secteur industriel. En termes d'impact scientifique, les articles auxquels les doctorants ont contribué obtiennent un nombre moyen de citations significativement plus faible celui des autres papiers québécois auxquels ils n'ont pas contribué, même si le facteur d'impact moyen de revues dans lesquelles ils publient est, au contraire, significativement plus élevé. Cela suggère que les articles des doctorants souffrent de l'effet St-Matthieu, phénomène selon lequel la paternité d'une découverte sera plus aisément attribuée à un chercheur reconnu qu'à un autre l'étant moins.

Dans l'ensemble, cette thèse interdisciplinaire fournit une mesure unique de la prévalence, du contexte et de l'effet de la socialisation à la recherche dans les programmes de doctorat ainsi qu'une meilleure compréhension de l'importance des doctorants au sein de la communauté scientifique québécoise. Ces résultats devraient être d'un grand intérêt pour les administrateurs d'université, les conseils subventionnaires ainsi que les chercheurs dans le domaine des politiques scientifiques.

Mots-clés: bibliométrie; scientométrie; doctorat; Ph.D.; étudiants; enseignement supérieur; universités; Québec; Canada.
Table of Contents

Abstract .................................................................................................................................... v

Sommaire ............................................................................................................................... vii

Table of Contents .................................................................................................................. ix

List of Tables ......................................................................................................................... xi

List of Figures ...................................................................................................................... xiii

List of Abbreviations ......................................................................................................... xvii

Acknowledgements ............................................................................................................. xix

1. Introduction ..................................................................................................................... 1

   1.1. Background ................................................................................................................. 3
       1.1.1. Students’ socialization to research ..................................................................... 3
       1.1.2. Systems of scientific research ........................................................................... 7

   1.2. Research questions ..................................................................................................... 9

2. Literature review .............................................................................................................. 15

   2.1 Students’ participation in the publication process ................................................ 15

   2.2. Factors affecting students’ publication activity .................................................... 26

   2.3. What (and who!) is a scientific author? ................................................................. 31

   2.4. Bibliometrics ............................................................................................................. 38
       2.4.1. Limitations of bibliometrics and differences among disciplines .................. 41

3. Research methods ........................................................................................................... 47

   3.1. Bibliometric database .............................................................................................. 50
       3.1.1. Disciplinary classification of journals ............................................................... 55
List of Tables

Table 1. Time lag between year of graduation and year of publication of dissertation-based publications (Anwar, 2004) ................................................................. 17

Table 2. Disciplinary classification of SSH journals (number of journals and papers presented in parenthesis), 2000-2007 ................................................................. 57

Table 3. Disciplinary classification of NSE journals (number of journals and papers presented in parenthesis), 2000-2007 ................................................................. 58

Table 4. Sample record of the GDEU database provided to the author ...................... 62

Table 5. Number and percentage of distinct given names and doctoral students to which a gender could be attributed, by method ......................................................... 64

Table 6. Number and percentage of doctoral students, by gender ............................. 65

Table 7. Classification of programs and number of doctoral students (in parenthesis), 2000-2007 ........................................................................................................ 67

Table 8. Sample record of the GDEU database, including modifications ................. 68

Table 9. Results of the matching of core papers at the levels of university researchers, articles and author-paper combinations ......................................................... 79

Table 10. Results of the matching of core papers and papers with the same keywords or cited references, at the levels of university researchers, articles and author-paper combinations .................................................................................................................. 83

Table 11. Results of the matching of core papers and papers with the same keywords or cited references, without the ‘same institution’ criteria, at the levels of university researchers, articles and author-paper combinations ............................................. 84

Table 12. Overall results of the automatic matching of papers, using core papers, keywords and references previously made, and the matrix of discipline of departments and specialty of papers .................................................................................. 89
Table 13. Percentage of faculty-students collaborations, by gender of doctoral students and of faculty, as a percentage of the total number of doctoral students' faculty-students collaborations, by discipline of the department, 2000-2007 .......... 144

Table 14. Percentage of Quebec and Canada’s papers, by sector, 2000-2007 .......... 156
List of Figures

Figure 1. Creation of the database of scientific papers to which doctoral students contributed .......................................................... 49

Figure 2. Potential (fictitious) homographs for John William Dawson ............... 50

Figure 3. Publications found in the WOS for John William Dawson .................. 50

Figure 4. Schematic representation of the bibliometric database.......................... 52

Figure 5. Percentage of papers assigned after manual validation, by researcher...... 75

Figure 6. First matching criteria for creation the core of papers............................ 77

Figure 7. Second matching criteria for creation the core of papers......................... 78

Figure 8. Percentage of rightly assigned and wrongly assigned papers, as a function of the keywords previously used by a university researcher, 2000-2003 and 2004-2007 81

Figure 9. Percentage of rightly assigned and wrongly assigned papers, as a function of the references previously made by a university researcher, 2000-2003 and 2004-2007 ......................................................................................................................... 82

Figure 10. Percentage of papers assigned after manual validation, for each combination of discipline of departments (x-axis) and specialty of publication (y-axis) ......................................................................................................................... 87

Figure 11. Number of rightly and wrongly assigned author-paper combinations, as a function of attribution percentage of papers from a discipline to authors from a department ......................................................................................................... 88

Figure 12. Interface for the manual attribution of publications to doctoral students 95

Figure 13. Percentage of doctoral students having published at least one paper, by first year enrolled in a doctoral program (cohort) and discipline of the doctoral program, 2000-2007 ......................................................................................................................... 114

Figure 14. Percentage of doctoral students and faculty members having published at least one paper, by discipline of the doctoral program, 2000-2007 ......................................................... 118
Figure 15. Scatter plot of the percentage of doctoral students with at least one paper, as a function of the number of faculty with at least one paper, by specialty of the doctoral program, 2000-2007. Inset: distribution of the ratio between the percentage of faculty publishing and doctoral students publishing, as a function of the percentage of doctoral students publishing................................................................. 121

Figure 16. Relationship between the average amount in funding received by professors and the percentage of doctoral students who published at least one paper, by specialty of doctoral programs, 2000-2007 (1999-2006 for funding).........................123

Figure 17. A) Distribution of the percentage of doctoral student having published at least one paper, by discipline of the doctoral program and gender, 2000-2007. B) Distribution of the difference between male and female PhD students participation in peer reviewed papers, by specialty of the department........................................125

Figure 18. Percentage of papers with more than one author (co-authorship) of PhD students’ papers and of other Quebec papers, by discipline of journal, 2000-2007.128

Figure 19. Average number of authors of PhD students’ papers and of other Quebec papers, by discipline of the journal, 2000-2007.................................................................128

Figure 20. Percentage of doctoral students’ papers written with faculty members or with other doctoral students from a Quebec university.............................130

Figure 21. Percentage of faculty papers authored with doctoral students from a Quebec university..................................................................................................................131

Figure 22. Average number of collaborating Quebec faculty in PhD students’ papers and average number of PhD students in Quebec faculty papers, by discipline of the department, 2000-2007.................................................................133

Figure 23. Co-authorship network of doctoral students and faculty members in A) laboratory medicine (more than 2 papers co-authored), B) Physics (more than 1 paper co-authored), C) Psychology (no threshold), D) education (no threshold), 2000-2007 ...............................................................................................................................135

Figure 24. Percentage of international collaboration of doctoral students’ paper and of other Quebec papers, by discipline of the journal, 2000-2007.........................136

Figure 25. Percentage of papers authored with researchers from sectors other than universities and hospitals, 2000-2007.................................................................138

Figure 26. Percentage of Quebec’s doctoral students’ papers that are first authored, by discipline of the journal, 2000-2007.................................................................139

Figure 27. Normalized author order of A) doctoral students and faculty papers and B) for papers co-authored by at least one doctoral student and one faculty member, by discipline of the doctoral program, 2000-2007.................................................................142
Figure 28. Average number of papers of doctoral students having completed their program and of doctoral students that have not completed their program as of the end of 2007, for 2000-2002 cohorts ............................................................... 147

Figure 29. Relationship between pre-graduation productivity and post-graduation productivity, for the subset of doctoral students who graduated in 2003-2004 (N=2,319) .................................................................................................................. 149

Figure 30. Quebec’s percentage of Canadian and World papers, 1980-2007 .......... 155

Figure 31. Average of relative citations of Quebec and Canadian papers, 1980-2007 ............................................................................................................................................... 157

Figure 32. Number of doctoral students and university researchers, by discipline of the department, 2000-2007 .................................................................................................................. 158

Figure 33. Percentage of Quebec university papers authored by doctoral students and faculty members, by discipline of the journal, 2000-2007 ......................................................... 161

Figure 34. A) Correlation between the percentage of Quebec papers co-authored and the percentage of Quebec university papers authored by doctoral students and B) Correlation between the average number of authors of Quebec papers and the percentage of Quebec university papers authored by doctoral students, both by specialty of the journal, 2000-2007 ............................................................................................................................................... 164

Figure 35. Percentage of Quebec university papers to which doctoral students contributed, by discipline of the journal (natural sciences and engineering), 2000-2007 ............................................................................................................................................... 166

Figure 36. Percentage of Quebec university papers to which doctoral students contributed, by discipline of the journal (social sciences and humanities), 2000-2007 ............................................................................................................................................... 167

Figure 37. Cumulative distribution of the number of papers of PhD students and of faculty, 2000-2007 .................................................................................................................. 170

Figure 38. Average number of papers of doctoral students and faculty members, by discipline of the program/department, 2000-2007 .................................................................................................................. 171

Figure 39. Average number of papers of doctoral students, by gender and discipline of the program, 2000-2007 .................................................................................................................. 173

Figure 40. A) Average of relative citations and B) Average of relative impact factor of papers to which doctoral students contributed and of other Quebec papers, by discipline of the journal, 2000-2007 ................................................................. 176
Figure 41. A) Average of relative citations and B) Average of relative impact factor of papers to which doctoral students contributed, either as first author or other author, by discipline of the journal, 2000-2007 ............................................................... 179

Figure 42. Percentage of doctoral students’ papers and of Quebec other papers among the top 5% most cited papers, by discipline of the journal, 2000-2007 ........ 180

Figure 43. Average of relative citations of papers to which doctoral students contributed, by gender and discipline of the program, 2000-2007 ................................. 182
List of Abbreviations

AHCI: Arts and Humanities Citation Index
ARC: Average of relative citations
ARIF: Average of relative impact factor
CAI: Commission d’accès à l’information du Québec (Quebec’ Information access commission)
CIHR: Canadian Institutes of Health Research
CIP: Classification of Instructional Programs
FQRNT: Fonds québécois de la recherche sur la nature et les technologies
FQRSC: Fonds québécois de la recherche sur la société et la culture
FRSQ: Fonds de la recherche en santé du Québec
GDEU: Gestion des données sur l'effectif universitaire (university enrolments data management)
MDEIE: Ministère du développement économique, de l’innovation et de l’exportation (Ministry for economic development, innovation and exportations)
NSE: Natural sciences and engineering
NSERC: Natural Sciences and Engineering Research Council of Canada
NSF: National Science Foundation (United States)
OST: Observatoire des sciences et des technologies
RRQ: Régie des rentes du Québec
SCIE: Science Citation Index Expanded
SNA: Social Network Analysis
SSCI: Social Sciences Citation Index
SSH: Social sciences and humanities
SSHRC: Social Sciences and Humanities Research Council of Canada
WOS: Web of Science
Acknowledgements

I was lucky to enter the world of research during my first year as an undergrad, before my twenties. I have, since then, been involved in several research projects, and have always wondered if I was the only student in this situation. This thesis started, thus, with a reflexive inquiry: to what extent are students in general—and PhD students in particular—involved in research?

My own integration into research has benefited from the help of more senior colleagues. First and foremost, I would like to thank my friends and colleagues Éric Archambault and Jean-Pierre Robitaille who, each in their own manner, shaped my understanding of the scientific community in general and of bibliometric methods in particular and encouraged me to pursue graduate studies. I would also like to thank my co-advisor Yves Gingras for teaching me with his rigorous—yet pragmatic—approach to research, my advisor Jamshid Beheshti for his support and encouragement, as well as the other members of my doctoral committee (France Bouthillier, John Leide and Pierre Pluye) for their careful reading if this thesis and numerous suggestions. Many, many thanks also to my colleague Jillian Tomm for her careful reading and re-reading of this thesis. Of course, any remaining errors and mistakes are entirely my responsibility.
At the Observatory des sciences et des technologies (OST-Université du Québec à Montréal), where this thesis was carried out, I would like to express my gratitude to Benoît Macaluso for his help on the conceptualization and implementation of the automatic attribution algorithm as well as to Mario Rouette for building the manual attribution interface. I would also like to thank all members and past members of OST, whose work on maintaining the bibliometric database made this thesis possible: Philippe Gendron, Pascal Lemelin, Jean Lebel, François Vallières, Sarah-Maude Gagnon, Valérie Bastien, François Taillefer, Yannik Melançon, Joanne Cocco, Lionel Vécrin, Guillaume Girard, Marie-Claude Laframboise, Louis-Michel Lebeau, Frédéric Mongrain, Cynthia Lisée, Pawel Dziob, Isabelle Massé, Nathalie Durocher, Jocelyne Raymond, Nicolas Bastien, Philippe Mirabel, Benoît Gagné and François Petit. I would also like to thank David Campbell, Grégoire Côté and Frédéric Trempe-Bertrand at Science-Metrix for interesting bibliometric discussions.

At the Centre interuniversitaire de recherche sur la science et la technologie (CIRST), I acknowledge the constant support of Lucie Comeau and Martine Foisy, as well as great discussions with Vincent Bontems, Jean-Philippe Gendron, Brigitte Gemme, Julien Prud’homme, Étienne Vignola-Gagné, Alain Couillard, Sébastien Mosbah-Natanson and Matthew Wallace. During my doctoral studies at McGill’s School of Information Studies, I have also enjoyed exchanging with several other doctoral students, among others Lorie Kloda, Charles-Antoine Julien and Rafa Absar. I have also enjoyed my debates with Dawn Allen and all the class of EDEM-692 about quantitative and qualitative methods. Special thanks also to my friend Steve Poirier for lengthy discussions on science in general.
Finally, I would also like to express my sincere gratitude to the Canada research chair in history and sociology of science (year 1), SSHRC Joseph-Armand Bombardier CGS Doctoral Scholarship program (years 2, 3 and 4) and the International Society for Scientometrics and Informetrics (Eugene Garfield doctoral dissertation scholarship award) for providing me with the funding essential to pursue this research. I would also like to thank the Quebec’s Ministère du développement économique, de l’innovation et de l’exportation, and especially Christian Villeneuve, as well as Quebec’s three research councils (Fonds de la recherche en santé du Québec, Fonds québécois de la recherche sur la nature et les technologies and Fonds québécois de la recherche sur la société et la culture) for providing the list of Quebec university researchers. I would also like to thank the Canadian G13 data exchange for their help in the disciplinary classification of departments and research programs. I would also thank Quebec’s Ministère de l’éducation, du loisir et du sport for providing me with the list of all doctoral students in the province.

Merci ultime à mes parents, mes sœurs et mes amis pour leur soutien indéfectible. À Noah et Zoé que, je l’espère, auront la chance de mener leurs propres recherches un jour. Mais seulement si ça leur tente. À Véronique qui m’a permis de vivre plus d’une vie à la fois.
1. Introduction

Graduate students are an important part of the academic workforce. In Canada, more than 32,000 students were enrolled in PhD programs for the year 2003, and more than 3,800 graduated the same year (Canadian Association for Graduate Studies, 2006). Of those students, about 30 % were studying in Quebec. There are almost as many doctoral students as full-time university professors in Canada—40,800 in 2006 (Association of Universities and Colleges of Canada, 2007). For the province of Quebec, the tendency is similar: about 27,500 distinct students were enrolled in doctoral programs at some point in the province between 2000 and 2007 (GDEU database, compilations performed by the author), while there were 9,306 university professors in the province in 2006 (CREPUQ, 2009).

During their apprenticeship, graduate students perform several activities related to both the formal and hidden curricula of the programs and disciplines in which they are studying. For instance, they participate in academic life by sharing research ideas with colleagues and faculty members during staff meetings, working as teaching and research assistants, etc. One other aspect of the hidden curriculum is participation in the publication of scholarly articles which, it is assumed in this thesis, is an indicator of their socialization and integration into research. Though graduate students have been the object of important scholarly attention—especially in the sociology of higher education—no study has attempted to measure the scientific publication
activity of an entire population of graduate students, the conditions under which they participate in research, and the importance of their contribution, as ‘would-be’ researchers, to the advancement of knowledge.

The purpose of this thesis is to investigate the scientific publication activity of the whole population of PhD students enrolled in Quebec’s universities. More specifically, this thesis has two research goals, to which several questions are associated. The first research goal of this thesis is to measure the extent to which graduate students participate in the publication process during their studies and, hence, are socialized and integrated into the research activity. In addition to the extent of this socializing practice, this thesis also investigates the nature of the context in which these papers are produced, as well as their relationship with the subsequent research careers of graduate students. The second research goal is to assess the contribution of graduate students to research within Quebec’s research system, which includes in this thesis all domains of scholarly activity (natural sciences, medicine, social sciences, arts and humanities).

This PhD thesis has six chapters—including the introduction and conclusion—and can be divided into two parts. The first section (chapters two and three) presents the literature relevant to this thesis and the methods that are used. More precisely, chapter two presents the literature on students’ contributions to research—as measured by their participation in peer-reviewed papers—as well as the literature on the factors that influence this participation. It also discusses the question of authorship, presents some of the basic concepts of the field of bibliometrics, and
provides a detailed overview of the limitations of bibliometrics analysis and differences in publication and citation practices across disciplines. Chapter three describes the methodological challenges for this study as well as the methods used. The second part of this thesis (chapters four and five) presents and analyzes the results. Chapter four presents the data in light of PhD students’ socialization to research, while chapter five presents data on their contribution to Quebec’s research system. Chapter six presents the general conclusions and limitations of this thesis, as well as further research to be conducted on the topic.

The remainder of this introductory chapter provides the background for this thesis and reviews the literature on the process of socializing students to research and on the ‘reproduction’ of the scientific community. It also introduces theoretical and empirical studies on the structure of the scientific community at the level of research institutions¹ and presents the detailed research questions that are answered in this thesis.

1.1. Background

1.1.1. Students’ socialization to research

During their studies, PhD students are in an intermediary position (Delamont, Atkinson and Parry, 1994). In addition to their traditional role of learners, the purpose of which is to acquire new knowledge and skills, it is also expected that they contribute to the advancement of knowledge in their scientific discipline. As mentioned by Ziman (1993, cited in Sadlak, 2004, p.8), “…the PhD experience is the

¹ In this thesis, the word “institution” refers to organizations that are involved in research (universities, government laboratories, industries, hospitals, etc.)
psychological transition from a state of being instructed on what is already known to
a state of personally discovering things that were not previously known.” This is,
ultimately, the purpose of doctoral programs: to form new researchers who can
contribute to the advancement of knowledge and, subsequently form other new
doctoral students. In other words, it is expected that they participate in academic life
in a manner similar to that of established researchers—at least by the time they
complete their program. As shown by Hagstrom (1965), graduate students make the
‘gift’ of their research results to other researchers by producing a thesis, which is
evaluated by researchers established in their field. In return, students receive
recognition from their peers, hence confirming their status as members of the
scientific field (Bourdieu 1975; 2001). In parallel with the thesis, with which they
traditionally gain access to the scientific community, they may also submit their work
to a scientific journal or to an academic publisher.

It is during this period of their lives that graduate students get socialized into
research; that they acquire the behaviours, attitudes, norms and know-how of their
scientific community. Though it is also during this period that they learn the
knowledge specific to the research field they are in, it has been revealed that this is
less important—in the scientist’s own view—than the savoir-être (the how to behave
like a scientist) that they learn during that apprenticeship (Zuckerman, 1977).
Similarly, Campbell (2003) uncovered, from a series of interviews with PhD students
in science, that what is formally taught to graduate students is not as important as
what they learn by interacting with more senior students and faculty members. In
other words, socialization to research can be understood as the process by which
young researchers acquire the research *habitus* (Bourdieu, 1980; 2001), that is, how to think and act like researchers. In the same line, Golde (1998) has defined the process occurring in graduate school as a double socialization: the socialization to being a *graduate student* and the socialization to being a *researcher*—the future career of a significant share of graduate students. Other authors such as Turner and Thompson (1993) and Gardner (2007) have also positively linked socialization to degree-completion.

Socialization to research is indeed very complex and includes a wide spectrum of activities, such as lab work, writing of research proposals, staff meetings, etc. One of these components is the publication activity, which can be defined as the process by which newly created scientific knowledge goes, through peer review, from the lab to the scientific community. Publication can be considered to be a very important component of the socialization and integration of students to research, because it is through that process that knowledge gets validated—or rejected—by the scientific community. Given the increased competition of academia today, the mere writing of a thesis, though necessary, cannot be considered as a sufficient condition to enter today’s scientific community. In other words, it is considered that the complete training of doctoral researchers does not end with the writing of the thesis, but includes the publication of its results in the scientific community. A consequence of this increased competition is that, though about half of all graduate students aim for a university career (Fox and Stephan, 2001), only a third of Quebec’s PhD graduates become integrated into faculty (Conseil Supérieur de l’Éducation, 2003). Along the same line, citations made to theses—the standard output of doctoral research—has
declined steadily since the mid–seventies (Larivière, Zuccala and Archambault, 2008). Hence, in order to be integrated into academia today, thesis writing has to be combined with the writing of research articles, which is an indisputably important component of that process.

Given that a central component of the research *habitus* is to publish new knowledge—in scientific papers or books—it could be considered that it is more by participating in that publication and diffusion effort that graduate students are socialized into research and integrated into the scientific community than by the sole writing of a thesis. Though students defend their theories/discoveries/ideas in their theses, which they submit to a group of peers (the doctoral committee), this exercise takes place in a ‘controlled atmosphere’, given that the committee’s composition is often decided with the collaboration of their advisors—the latter having an interest in seeing their students graduate. Hence, the thesis submission and defence can be considered more as a *rite of passage* (Bourdieu, 1982; Goffman, 1974), as opposed to standard peer review that is experienced when submitting a paper to a journal. When graduate students submit papers to scientific journals, they are in the real scientific world and have no control over the choice of their evaluators. It is this form of peer review that will become the rule if they pursue a career in research after they graduate.

Analyzing doctoral students participation in the publication process can shed light on 1) the extent to which graduate students’—as a social group—are socialized and integrated into research, 2) the context (collaboration or sole authorship,
collaboration with industries, etc.) in which this process occurs and 3) the impact it has on students’ research careers.

1.1.2. Systems of scientific research

Several theoretical frameworks have described how knowledge production is organized at the level of research organizations. One of the most well known is the National Innovation System, which could be defined as the “set of institutions whose interactions determine the innovative performance of national firms” (Nelson, 1993, p.4, cited in Godin, 2009). As argued by Niosi (2000 and 2005), research in Canada is structured in a ‘system’ in which universities, governments and industries collaborate to create new knowledge, innovate and, ultimately, commercialize products, processes and services. Similarly, Gibbons et al. (1994) make a case for a ‘new’ system of knowledge production, in which, among other things, knowledge is increasingly produced in collaboration between universities and industries (mode 2), instead of by universities alone (mode 1). In this new research system (mode 2), the research problems are more applied and multidisciplinary, and the organization of research is less hierarchical. Along the same lines, Leydesdorff and Etzkowitz (1996) argue that knowledge production is structured in a ‘triple helix’, where universities, governments and industries collaborate in various manners.

Though these frameworks are of interest for studying the relations between organizations involved in the knowledge production process, they give few clues on the dynamics of the research system at a more micro level. For example, Gibbons et al. (1994) suggest that universities are becoming less central as knowledge producers.
Quantitative studies, however, have shown for the Canadian case that the proportion of papers from universities has increased steadily since the beginning of the 1980s (Godin, Doré and Larivière, 2002; Godin and Gingras, 2000a). Still, none of these studies provide any insight on the status of the authors of these papers and, hence, on the specific contribution to research of each social groups within these organizations. In other words, we do not know, in the case of universities, to what extent graduate students, post-doctoral fellows and technicians are contributing to research, and generally assume that publications from universities are written by faculty members. These are empirical questions that can be answered only with empirical data; this study provides insight into the extent of the contribution made by PhD students to Quebec’s research system, as measured by their participation in peer-reviewed papers produced by the institutions active in the system.

In this thesis, Quebec’s research system is defined as the sum of all scholarly research conducted in the province, be it in the natural sciences, medicine, the social sciences, the arts or the humanities. In other words, the research system of the province consists of the sum of all papers authored by researchers affiliated with Quebec organizations, irrespective of the discipline. Hence, contrary to Nelson (1993) and Lundvall (1992), we do not restrict ourselves to economically useful knowledge or to innovation, but include all fields of scholarly activity in which doctoral degrees are granted.
1.2. Research questions

Two general research questions are answered in this thesis, each of which corresponds to a distinct research goal. The purpose of the first research question is to investigate one component of graduate students’ socialization and integration into research, that is, their participation in peer-reviewed publications. As mentioned earlier, socialization into research is a complex process that has several components, one of them being participation in the publication of peer-reviewed papers. Three aspects of this socialization are studied: the extent of this socialization, its context and its impact.

First, this thesis determines the extent to which Quebec’s PhD students, as a social group, are socialized into research, i.e., the higher the share of PhD students that have participated in the publication process, the more we can assume that students are actually integrated and socialized into research. This integration into research is analyzed over time, by discipline of the doctoral program, number of years in the program, gender, etc. By extension, the extent to which socialization and integration into research are important components of the hidden curriculum of PhD students is also measured. Second, the context of this socialization is studied using co-authorship relations (student–faculty, student–student, etc.) as well as interinstitutional collaborations (university–university, university–industry, collaboration with foreign research organizations, etc.). Thirdly, the impact of socialization on the subsequent research career of graduate students is assessed by measuring the relationship between publication and degree completion, as well as between pre-graduation publication and post-graduation research productivity.
As a second research question, this thesis assesses the importance of students in the system of knowledge production. The authorship of a scientific paper attributes to the author(s) the credit (and responsibility) for the research results published (Birnholtz, 2006; Pontille, 2004). Thus, by measuring papers authored and co-authored by students we can get a good measure of their overall contribution to the research system. This thesis uses data on all Quebec doctoral students and can, thus, be considered a large-scale study of a complete population. More specifically, this thesis answers the following research questions:

1. **Primary research question:** What is the extent of doctoral students’ socialization and integration to research, its context and its impact on their careers?

   1.1. What is the share of Quebec’s doctoral students who participate in scientific publications during their PhD studies? Is this share changing over time? Are there differences among disciplines and gender?

   1.2. What is the context of this participation in knowledge production? Given that research is often a collaborative enterprise (see, e.g., Larivière, Gingras and Archambault (2006) for the Canadian situation and Larivière (2007) for that of Quebec), are students’ papers more likely to be the effort of collaboration (supervisor/student, students/students, etc.)?

   1.2.1. Is the funding received by faculty members having an effect on doctoral students’ participation in papers?
1.2.2. Given the changes in the nature of academic research (Gibbons et al., 1994; Leydesdorff and Etzkowitz, 1996), are students co-publishing more with non-university organizations, such as industries?

1.2.3. Are students' publishing patterns (international collaboration, collaboration with industries, etc.) different from those of established researchers?

1.2.4. In cases where students are co-authors, are they more likely to be first or last authors?

1.2.5. Following Neumark and Gardecki (1998), are female students more likely to work with female university-based researchers and professors?

1.3. What is the effect of this socialization on the ulterior research careers of doctoral students?

1.3.1. Given that a significant part of PhD students do not complete their degrees (Gardner, 2007), is participation in the publications process affecting their degree completion as well as time to completion?

1.4. Is the participation in peer-reviewed papers affecting their research careers as well as their subsequent productivity?

2. Secondary research question: What is the contribution of PhD students to Quebec’s research system?

2.1. What is the percentage of Quebec scientific publications to which graduate students have contributed?
2.1.1. Are there differences among disciplines and gender?

2.1.2. Is this percentage changing over the period under study?

2.2. What is the average productivity of doctoral students?

2.2.1. Are there differences among disciplines and gender?

2.2.2. How does doctoral students’ productivity compare with that of established researchers?

2.3. Do scientific papers to which doctoral students have contributed obtain a higher scientific impact—as measured by citations received—than those to which they did not contribute?

2.3.1. Is scientific impact influenced by the student’s place in the author order?

2.3.2. Is scientific impact influenced by the student’s gender?

In particular, the proportion of papers with students as authors or co-authors within all of Quebec’s scientific production is examined, as well as signature patterns, the order of authors, co-authorship with supervisors, collaboration between students, and co-authorship with international researchers, or with industrial or governmental partners. The post-graduation scientific output of students who graduated during the studied period is also analyzed. This allows exploring the relationship between doctoral students’ involvement in the publication process during the course of their studies and the pursuit of a career in research. In order to see if students’ scientific
papers have a high impact on science, an analysis of the number of citations received by these papers is also carried out.

These measures are broken down by students’ discipline and gender to determine if these two factors influence doctoral students’ contribution to published scientific literature. Finally, this study covers all disciplines of the natural sciences, medicine, social sciences, arts and the humanities over the 2000-2007 period.
2. Literature review

This chapter reviews the relevant literature on the participation of doctoral students to scientific publications as well as the factors that influence this participation. It also examines the question of authorship and the criteria used in the various disciplines to grant it. Finally, the last section of the chapter presents the field of bibliometrics and provides a detailed overview of the limitations of bibliometrics and of the differences in publication and citation practices across disciplines.

2.1 Students’ participation in the publication process

Despite the fact that the participation of graduate students in the publishing process is important to the understanding of 1) the extent to which students get socialized to the research activity and 2) how significant students’ scientific contributions are to the research system, very few studies have attempted to measure graduate students’ participation in the production of knowledge. As Nettles and Millett (2006) put it, “[a]lthough students are believed to acquire preparation for their lives as scholars and researchers while attending graduate school, evidence of their research productivity during their doctoral students days is not abundant. [...] The dissertation [...] is the only research product for which there is comprehensive documentation” (p.105).

This section reviews the literature on students’ contribution to research, as measured by their contribution to published scientific literature.
From the qualitative analysis of a physics laboratory, Shinn (1988) showed that the research results of junior researchers had a greater cognitive value than those of senior researchers and, because of their greater precision, were often able to end scientific controversies. Also studying experimental physics, Walford (1983) came to the conclusion that students make “a significant contribution to research” (p.253), though these contributions are often of a technical nature. Nevertheless, no study has yet attempted to measure that contribution quantitatively at the macro level. The main reason for that is the technical difficulty associated with the identification of authors and their status (professor, student, postdoctoral researcher, etc.). Treatment of this difficulty is discussed in detail in the research methods section.

Because of these technical limits, most studies conducted so far are micro-level case studies focusing either on a small sample of students or on specific fields. Also, given the difficult task of obtaining an assessment that is objective and independent of the actors’ own opinions of their contributions—advisors and students often disagree on their respective roles in student research (Berelson, 1960; Campbell, 2003)—most of these studies were made using bibliometric methods. For the field of information sciences, Anwar (2004) examined the pre-doctoral (1991-1995) and post-doctoral (1996-2000) publication activity of 54 individuals who graduated from U.S. universities in 1995. His data provided evidence that 24 graduate students out of 54 (44.4%) published one document or more (journal article, conference paper, book or annual review) during the course of their PhD, while a third of all graduates did not publish anything over the ten-year period studied. Another interesting feature of his study is the analysis of publications that were derived from dissertations, which
showed that 50% of the students wrote at least one publication from their dissertation research. This allowed the author to compile *publication lags*, that is, the time elapsed between the year of graduation and the publication of a paper derived from the dissertation. As presented in Table 1, most papers produced as the result of a dissertation were published within two years of graduation, which indicates that research findings drawn from the thesis can still be published a few years after the completion of the degree². Finally, Anwar’s data provided evidence of the very low level of collaboration between students and advisors, only one paper out of his sample being the result of such collaboration.

Table 1. Time lag between year of graduation and year of publication of dissertation-based publications (Anwar, 2004)

<table>
<thead>
<tr>
<th>Year of publication</th>
<th>Lag (years)</th>
<th>N. Pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>-3</td>
<td>2</td>
</tr>
<tr>
<td>1993</td>
<td>-2</td>
<td>2</td>
</tr>
<tr>
<td>1994</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1995</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1996</td>
<td>+1</td>
<td>14</td>
</tr>
<tr>
<td>1997</td>
<td>+2</td>
<td>14</td>
</tr>
<tr>
<td>1998</td>
<td>+3</td>
<td>5</td>
</tr>
<tr>
<td>1999</td>
<td>+4</td>
<td>4</td>
</tr>
<tr>
<td>2000</td>
<td>+5</td>
<td>2</td>
</tr>
</tbody>
</table>

For the field of adult education, Blunt and Lee (1994) used survey data to study the contribution of students to papers published in the *Journal of Adult Education/Adult Education Quarterly* over the 1969–1988 period. Their data shows that 113 students

² This table has methodological implications for this thesis, as it will help decide how long after graduation papers will be considered as “graduate students” papers and not as post-Ph.D. papers.
contributed—either as authors or co-authors—to 128 articles published in the journal, accounting for a large proportion (46%) of all papers published in the journal. Male students accounted for 69% of all students’ papers, while females participated in 31% of the papers—though at an increasing rate over the period studied. The authors also analyzed collaboration trends of students: 55% of the papers were written by one student alone, 39% had two authors and 6% had more than two authors. These collaborators were either students or non-students.

In the field of medicine, Cursiefen and Altunbas (1998) examined the research output of medical students of a German medical faculty over the 1993–1995 period. Using Medline, they found that students were the authors of 28% of the entire faculty’s papers (316 out of 1128) and the first authors in slightly less than half of the cases. Similarly, Whitley, Oddi and Terrell (1998) studied students’ participation in the publication process for the field of nursing, surveying the 633 authors publishing in the journal Nursing Research between 1987 and 1991. Their data showed that 31.6% of all authors publishing in that journal for the period were students. Though their survey did not include any information on students’ formal collaboration with their supervisors for the paper(s) they published in the journal Nursing Research, 84.5% of the students surveyed indicated that they were supervised by Faculty during the process.

In an attempt to provide data for a spectrum of subfields, Lee (2000) analyzed the participation in papers for a sample of PhD students in analytical chemistry, experimental psychology and American literature. Using Dissertation Abstracts to build
a sample of PhD students, he then compiled publication data for seven cohorts of students graduating between 1965 and 1995. In order to consider only publications written during PhD years—but also to take into account publications delays—Lee counted publications three years prior to graduation, and one year after graduation. Hence, for a student graduating in 1970, Lee considered papers published between 1967 and 1971, which is one year less than what was considered by Anwar (2004).

As might be expected, Lee’s data presented interesting differences among the three fields: while about 85% of 1995 graduates in analytical chemistry published at least one paper during their studies, this percentage dropped to 50% in experimental psychology and 35% in American literature. Lee also shows that, at least for the fields of analytical chemistry and American literature, there was an increase in student participation in the publication process over the period. In terms of collaboration trends, the author observed a decline in solo authorship in all three fields similar to what is observed at the macro-level.

For chemistry and physics, using a small sample of graduates (138 in physics and 237 in chemistry), Nederhof and van Raan (1987 and 1989) compared the productivity and citation rates of *cum laude* (with distinction) doctorate holders with that of *ordinary* doctorate holders, both during the course of their PhD and after graduation. Unsurprisingly, they point out that *cum laude* graduates had both higher productivity and citation rates before graduation than non *cum laude* graduates, which suggests that judgments of senior scientists of students’ quality are well correlated with bibliometric measures of productivity and scientific impact. Though both samples
revealed that productivity of *cum laude* graduates was still higher than that of non *cum laude* after graduation, citations rates of both groups no longer differ significantly.

Apart from the micro-level studies described above, two comprehensive studies have been conducted on the topic to date. The first is by Liang, Liu and Rousseau (2004), who analyzed data from the Chinese Science Citation Database (CSCD)—a database of Chinese scientific periodicals similar to the Science Citation Index (SCI)—and identified publications by graduate students enrolled in three Chinese universities (Beijing University, Tsinghua University and Beijing Medical University) by merging the CSCD with the Chinese Dissertation Document Bibliography Database (CDDDB)\(^3\). In addition to standard bibliographical information, this dissertation database contains personal information about authors such as their age, gender, degree (master’s or doctorate) and supervisor’s name.

Querying the CSCD, they found 13,373 papers co-published by graduate candidates and supervisors for the 1989–1998 period\(^4\). They also labelled—using keywords—papers as related or unrelated to the students’ dissertations. The authors then put forward a very simple indicator—called the g-ratio—that is calculated by dividing the number of graduate students’ first authored publications by the total number of co-written papers. A g-ratio greater than 0.5 means that more papers are first authored by graduate students, and vice-versa. The data thus obtained showed that, on average, papers co-written with doctoral students have a ratio of more than 0.8, while

---

\(^3\) This method is similar to the method used in this thesis.

\(^4\) The authors do not mention in the paper whether they removed articles written by homographs of either the graduate student or the supervisor which, in turn, could add some noise to their data.
for master’s students it is of less than 0.7. As might be expected, these figures are higher when published papers are related to students’ dissertations. The data showed that students are more often first authors of the papers they write with their supervisors, which either means that 1) students are the main contributors of the papers they co-author or 2) there is a lot of noblesse oblige (Zuckerman, 1968) in the Chinese system. However, their paper does not provide any data on students’ share of the Chinese research output, nor on the percentage of students that publish papers during their studies.

The second large-scale study is that of Nettles and Millett (2006), who, in 1996, surveyed more than 9,000 doctoral students in the U.S. who had completed at least one year of study in their PhD program. This major survey analyzed several dimensions of the PhD experience—financing, socialization, satisfaction, etc—among which scientific productivity is one component. Their results show that about one graduate student in two had published some of their research—be it as a conference paper, an article, a book chapter or a book—during their studies. More specifically, 66% of engineering students, 57% of humanities students, 52% of science and mathematics students, 47% of social science students and 40% of education students had some research output. This data is not, however, broken down by both discipline and document type; the high figures—especially in the humanities—could be caused by the inclusion of conference papers and posters.

---

5 Noblesse oblige is the act, by senior researchers, of giving more visibility to junior colleagues, generally by placing their own names last on research papers.

6 It is also not clear in the book what is meant by “conference papers”: are presentations without published abstracts/proceedings included or is the definition restricted to published conference papers? There is a major difference between the two, given the fact that
which can be considered as more ‘entry-level’ publications than journal articles or book chapters.

For instance, if only journal articles are used to measure doctoral students’ research output, 47% of engineering students and 44% of science students had some scientific production to their name, while only 22% of students in the social sciences, 19% of students in the humanities and 15% of students in education did so. Nettles and Millett’s data also showed that there were some variations between fields and students’ cultural background (the authors use race as an indicator) and that research productivity was positively linked with degree completion and time spent in the program, i.e., students with a publication record did not spend more time in the program. Moreover, the authors show that, in all fields but education and social sciences, men were, on average, publishing more papers than women. Though this study shows that socialization into research is more important in the natural sciences and engineering (NSE) than in the social sciences and humanities (SSH), it does not provide any clues regarding the proportion of papers in the research system to which students participated or any evidence about the scientific impact of this research. This research also does not show the impact of scientific production on young researchers’ careers.

Gemme and Gingras (2008) conducted a similar survey, albeit on a smaller scale (104 respondents), on the socialization into research of graduate students (at the master’s

published conference papers usually go through the standard refereeing process, while unpublished conference presentations are usually accepted/rejected on the basis of an abstract.
and PhD level) in Quebec. Though their dataset is not large enough to compile meaningful statistics on publication practices of students by field or level of study, their data nonetheless show that slightly more than 80% of students had contributed to at least one publication since the beginning of their graduate program. Indeed, 55% of students had contributed to at least one conference paper, 43% to at least one research report, 41% to at least one article and 39% to at least one poster. Very few students contributed to books (2%) or book chapters (5%).

Another body of relevant literature concerns the relationship between the age of researchers and their scientific productivity or scientific impact; both measures often considered as indicators of creativity (Simonton, 1997; Dietrich and Srinivasan, 2007). This literature is relevant, given that PhD student are in general in their twenties and early thirties and hence can be considered as young researchers. One of the first studies on the topic was performed by Lehman (1953) and showed that the significant scientific contributions of researchers were generally made before their forties. Since this seminal paper, both sociologists and psychologists of science have extensively studies of the relationship between aging and research productivity (for a review of the topic, see Feist, 2006). These studies can be divided into two groups reporting opposite findings.

The first group of studies is based on extraordinary achievements—like Nobel Prizes—and find that these tended to occur before the age of 40 (Adams, 1946; Zuckerman and Merton, 1973; Zuckerman, 1977; Stern, 1978; Simonton, 1994; Dietrich and Srinivasan, 2007). We can also categorize in this group several studies
on the productivity of scientists which find that younger researchers are more productive than older ones (Gieryn, 1981; Horner, Rushton and Vernon, 1986; Over, 1988).

The second group of—generally more recent—studies on the topic report different results: it is not the younger researchers, but it is the ones in the middle of their careers and older, who produce more research and have a higher scientific impact (Dennis, 1956; Allison and Steward, 1974; Cole, 1979; Wray 2003 and 2004; Kyvik and Olsen, 2008). These two groups of studies have provided empirical evidence for two different theoretical explanations of the link between age and research productivity or scientific impact. The first group of studies, showing that young researchers publish more papers and are more likely to publish high impact papers, support Simonton’s model of creativity (Simonton 1984 and 1997), according to which each individual has an initial ‘creative potential’, and that this potential decreases over time. Similarly, these results also make sense in view of Kuhn’s (1962) argument that young researchers bring a fresh view to scientific problem—they do not yet have a strong commitment to the dominant paradigm of their disciplines—and, hence, are more likely to cause scientific revolutions.

The second group of studies, which reveals that older researchers are more productive and have more impact, supports Merton’s (1968; 1973) sociological analysis of the scientific community. Indeed, Merton shows that the scientific community is a gerontocracy; age being an important component in the stratification system of science (Cole and Cole, 1973). Indeed, as scientists’ age, they rise in the
academic hierarchy, obtain more funding, have more graduate students, gain access
to more resources and, hence, increase their scientific productivity and research
impact.

Along these lines, Gingras et al. (2008), using a large population—all professors and
university affiliated researchers in Quebec—provides evidence that both theories are
right from their relative perspectives, as one has to distinguish between scientific
productivity and impact. Indeed, their data provides evidence that, although older
professors publish more papers—they are the leaders of their own teams and have
access to more and better resources—, younger researchers’ papers have a higher
scientific impact, as measured by citations received. Given that this study is one of
the few large-scale attempts at compiling data at the level of individual researchers,
and that the author of this thesis has been heavily involved in this project, the
methods it used are analyzed in depth in chapter 3.

Finally, another study worth mentioning is that of Larivière, Zucca and
Archambault (2008) on the formal scientific impact of doctoral research results. The
question behind this research is whether the PhD thesis is an effective form of
diffusion for doctoral research results. The authors analyzed the relative importance,
among all cited material, of citations made to theses over the 1900–2004 period.
Their data point out that, since the beginning of the seventies, PhD theses account
for a smaller and smaller proportion of all cited literature. Though theses were more
cited in the SSH than in the NSE, their proportion among all cited material
decreased in all fields. Although one could argue from this study that PhD research
projects provide results that are less and less relevant for the scientific community, the authors suggest that this decrease is due to the fact that 1) in the natural sciences, engineering and medical fields, theses are increasingly written in the form of a series of articles (Breimer, 2010; Holdaway, 1994), and 2) in the social sciences and humanities, theses are generally subsequently published as books or articles. Hence, for obvious reasons of time, researchers prefer to read and cite the material derived from a thesis rather than from the thesis itself.

In sum, these mostly micro-level studies only provide a partial view of student participation in the publication process. Though they all follow a valid method—survey or bibliometrics—they are generally too narrow in scope (one or a couple of subfields) and do not provide any clues on the evolution of the phenomenon over time. In fact, even if we were to combine these studies, which we cannot given the different methods used in each study, we would not obtain a general portrait, since subfield studies only represent a small fraction of the scientific world. These studies neither provide the macro-level context of the subfield they study or go beyond the descriptive point. For instance, they do not discuss whether students’ papers are different from faculty papers, etc. Taken individually, these studies nonetheless tend to indicate that students do contribute to scientific publications and that, in some fields, their contribution is substantial.

2.2. Factors affecting students’ publication activity

Several factors can influence doctoral students’ participation in papers. The general literature on students’ participation in papers being scarce, only a few direct sources
of information on the specific factors influencing such participation were found in
the literature.

The main source of information on the topic found in the literature is the very large-
scale survey of doctoral students—more than 9,000 students covering all
disciplines—performed by Nettles and Millets (2006), which provides unique results
on the effect of funding and supervisors on students’ participation. The authors
present the different types of funding offered to students during their doctorate,
broken down into three categories: fellowships, research and teaching assistantships.
Unsurprisingly, students of the different disciplines are not equal in terms of access
to this research funding. More specifically, although 69% of students in humanities
received fellowships, this percentage was 61% in social sciences, 59% in sciences and
mathematics, 50% in the engineering and 46% in education. Nettles and Millett’s
study does not provide any indication of the amount received, as smaller fellowships
might explain the high percentage of students funded in the humanities.

In terms of research assistantships, the tendency is quite different: 82% and 69%,
respectively, of students in engineering and in science and mathematics worked as
faculty members’ assistants, while this percentage was only 49%, 33% and 28% in the
social sciences, humanities and education, respectively. Finally, in all disciplines but
education, a majority of students received teaching assistantships. These various
types of funding were found to have a strong effect on the PhD students’
participation in peer-reviewed papers. Students receiving fellowships were more
likely to publish in education, science/mathematics and in social sciences. Similarly,
in all disciplines but the humanities, students who were research assistants published more papers. Teaching assistantship was positively linked with research productivity only in the humanities. Along the same lines, Buchmueller, Dominitz and Hansen (1999) have also positively linked students’ research assistantships with research productivity, using a sample of doctoral students in economics. Unsurprisingly, working with productive faculty members also increased students’ research output.

Nettles and Millets (2006) analysis of faculty-student interaction revealed that these interactions were more frequent for students who expect their first job to be a faculty or postdoctoral position. Similarly, those who declared having mentors also had more faculty-students interactions. Mentoring was also found to have a strong effect on research productivity. This positive effect—students with mentors publishing more papers—was observed in all disciplines.

Paglis, Green and Bauer (2006) also analyzed the relationship between mentoring and research productivity for 161 ‘hard’ science doctoral students. Although their research is based on ‘submitted’ papers instead of papers actually published—which is an obvious limitation—, they found that ‘collaborative’ mentoring—defined as a mentoring relationship in which the students are asked by their mentors to be the co-authors of papers, conference papers, books, book-chapters or research grant proposals—was significantly associated with research productivity. Basically, this means that co-authorship with advisors/mentors had an influence on students’ publication activities.
In a survey of doctoral students in Quebec, Gemme and Gingras (2008) analyzed the importance of supervisors in the selection of their research problem. Although doctoral theses’ research topics are not necessarily those of the papers doctoral students work on, this is very likely related. Unsurprisingly, supervisors of more than 92% of respondents had been involved in the definition of the students’ research projects. This percentage was below average in pure sciences and mathematics (85.7%), social sciences and humanities (88.2%) and health sciences (91.7), but above average in applied sciences and engineering (95.1%). Co-supervisors were also involved in 62.5% of case (all disciplines combined), although this percentage was significantly lower in the social sciences and the humanities (35.3%). In these disciplines, however, ‘other’ non-supervising professors are more likely to be involved in the definition of the research. Students in the sciences were also more likely to choose their supervisors before their research subject, while in the social sciences and the humanities the opposite is observed: students choose their research subject, and then find a supervisor who is interested to supervise a thesis on the topic.

Similar results for the domains of natural and social sciences were also observed by Delamont, Atkinson and Parry (1997) and Ridding (1996): Although supervisors were very active in the choice of research topics in the sciences, students in the social sciences were more independent in that respect, which is likely a consequence of the more important collaborative nature of the former group of disciplines. Different disciplines have indeed different manners of enrolling doctoral students into research teams, and doctoral students in the natural or medical sciences are more likely to be
part of research teams than their colleagues of the social sciences. As pointed out by Pole et al. (1997), “In team-based research the PhD student may have a role as a formal member of the team which means that he/she contributes to a sizeable research project which is likely to make a significant contribution to the discipline in which it is based” (p. 58). These links between the work of the supervisors and that of the students might, in turn, increase their students’ participation in peer-reviewed papers.

Finally, these different ‘supervisory’ practices also have a strong influence on degree completion and on doctoral students’ satisfaction of their graduate school experience. As shown by Smeby (2000), although 73% of students in the natural sciences had completed their degrees by the end of their seventh year, less than half of students in the humanities and social sciences had done the same. Again the participation of students in research teams was noted as a determining factor: while students in the ‘hard’ sciences were directly involved in their supervisors’ research, those of the ‘softer’ social sciences and the humanities typically work alone. Similarly, as mentioned by Walford (1981) “The proportion [of doctoral students dissatisfied with the supervision that they receive] is generally found to be higher in the social sciences and arts than in the pure sciences, probably due to the much closer nature of supervision generally offered in the experimental sciences.” (p. 147-148)

This participation in research groups benefits the student—as they have more opportunity to publish and to graduate—but also the advisors, as those who supervise graduate students on projects related to their own research were more
productive than those who did not (Kyvik and Smeby, 1994). This effect was observed in the natural and medical sciences, but not in the social sciences and humanities. This can be explained by Kyvik and Smeby’s (1994) and others’ (Delamont, Atkinson and Parry; 1997; Pole et al., 1997; Ridding, 1996) observation that doctoral students’ research was more likely to be linked to that of their supervisors in the sciences than in the social sciences, or by the fact that co-authorship is less important in the latter group of disciplines than in the former (Lariviere et al., 2006).

2.3. What (and who!) is a scientific author?

Bibliometric methods are based on formal scientific contributions (Borgman, 1990), that is, contributions that are made public via published scientific documents. Hence, the understanding of who signs the publications and why they do so is of central importance because, in essence, the bibliometric unit of measurement is authorship. Though the notion of authorship has been largely neglected by the bibliometrics community, in recent years several researchers have discussed what it now means to be an author, mainly because of the increasing number of co-authored papers. Indeed, as shown by Larivière (2007), the number of papers co-authored in all fields of science, social sciences and humanities increased from about 10% of all published literature in 1900 to more than 70% in 2005. In the NSE it is even higher: more than 90% of all papers are co-authored (Larivière, Gingras and Archambault, 2006).

Given the changes in collaboration patterns, many studies on authorship are made from an insiders’ point of view—especially in medicine —being aimed at defining
what an author is in order to provide guidelines for research collaborations (Engers et al., 1999; Parker and Berman, 1998; Rennie, Yank and Emanuel, 1997; International Committee of Medical Journal Editors 2006). Some researchers in the sciences even discuss guidelines for including students as authors (Costa and Gatz, 1992; Fine and Kurdek, 1993; Oberlander and Spencer, 2006). However important these papers are to help solve controversies internal to the scientific community on authorship attribution, their aim is to prescribe norms of behaviour rather than understanding the phenomenon. After Foucault (1994), only recently did sociologists and historians discuss the meaning of authorship, how it has changed, and the differences among fields of knowledge (Biagioli, 2003; Birnholtz, 2006; Galison, 2003; Pontille, 2004).

In the first section of his paper, Birnholtz (2006) presents three functions of authorship in the scientific world. First, it attributes the credit for discoveries and ideas, second, it assigns ownership for them and, third, it enables the existence of an economy of reputation. The first and third functions are strongly linked, given that it is the attribution of this credit (or symbolic capital) that allows the existence of this economy of reputation. As extensively discussed by Bourdieu (1975; 2001), major discoveries recognized by Nobel Prizes do not necessarily generate patents or copyrights. The value of the scientific credit of discoveries is not money, but recognition by peers, reputation, etc. In other words, scientists are not rewarded with economic, but with symbolic capital, which they obtain through their publications.\footnote{This symbolic capital can, however, be transformed into other type of capital, such as economic capital (tenure, promotions, etc.) and social capital (relations).}

Likewise, as shown by Biagioli (1998), and in line with Merton’s (1973) norm of communism, scientific authorship differs from other forms of authorship (literary,
inventor of patents, etc.) in the sense that it is not linked with intellectual property or copyright. Hence, the notion of ownership in science has to be understood in terms of moral rights and responsibility. However, as illustrated by Kevles (1998), Biagioli (1999) and Wray (2006), collaborative research makes it more difficult to hold individual scientists accountable, especially in cases of scientific fraud. In sum, authorship provides credit to authors for their discoveries, and, in exchange, assigns responsibility for the reliability of the findings (Biagioli 1999).

In a comparative study of medicine and high-energy physics (HEP)—two fields with very high co-authorship rates—, Biagioli (2003) revealed that the meaning of authorship differs substantially among fields of knowledge. Hence, the attribution of signatures is also quite different. Using the case of the Collider Detector at Fermilab (CDF), he exposes the very clear and bureaucratic rules of authorship rules in high-energy physics: there is a standard author list, updated twice a year, on which 1) researchers with a PhD are included if they devote 50% of their time to an experiment, 2) graduate students are included if they work full time on an experiment, and 3) technicians are included if they make major contributions8 to the experiment.9 Those who leave an experiment remain authors of resulting papers for a year after they leave. In this field, authorship is thus seen in terms of “credits for accumulated labor” (Biagioli, 2003: 270), and we talk about contributors rather than authors. Also, given that everybody taking part in the experiment gets to sign the paper in alphabetical order, the attribution of scientific credit works differently in

8 Unfortunately, Biagioli does not define what is meant by “major contribution”.
9 Galison’s (2003) study of the Stanford Linear Detector (SLD) showed the existence of similar guidelines.
HEP. Biagioli concludes that this goes through colleagues’ appreciation, by letters of recommendation or personal communications. Birnholtz (2006) comes to similar conclusions.

In medical fields, the rules are much more blurred. In addition to being seen as a contributor, the author is also considered as a guarantor of the knowledge claims made in the paper, given that in these fields academic misconduct and fraud can have disastrous human impacts. Collaboration often occurs between labs—instead of within one large lab/experiment like in the HEP—and it is therefore more difficult to enforce clear guidelines. Though the International Committee of Medical Journal Editors (2006) established requirements for authors, the reality is much more complex, and the guidelines are not always followed (Bennett and Taylor, 2003). The guidelines state that:

“Authorship credit should be based on 1) substantial contribution to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published. Authors should meet conditions 1, 2, and 3.

When a large, multi-center group has conducted the work, the group should identify the individuals who accept direct responsibility for the manuscript (3). These individuals should fully meet the criteria for authorship defined above and editors will ask these individuals to complete journal-specific author and conflict of interest disclosure forms. When submitting a group
author manuscript, the corresponding author should clearly indicate the preferred citation and should clearly identify all individual authors as well as the group name. Journals will generally list other members of the group in the acknowledgements. The National Library of Medicine indexes the group name and the names of individuals the group has identified as being directly responsible for the manuscript.” (International Committee of Medical Journal Editors, 2006)

It should also be noted that the guidelines exclude from authorship technical help, writing assistance and general and monetary support (e.g., department chair, etc.).

Despite the existence of these guidelines, several problems remain in medical fields with regard to the assignment of authorship credit. Some articles have honorary, guest or gift authors—individuals that appear as authors but who do not meet authorship criteria—or ghost authors—individuals or organizations—such as industries—that should appear as authors but do not (Sismondo, 2009). Flanagin et al. (1998) measured the prevalence of both cases for a sample of 809 papers published in six medical journals in 1996. They found evidence of honorary authors in 156 articles (19%) and evidence of ghost authors in 93 articles (11%). Thirteen articles (2%) had both honorary and ghost authors. Unsurprisingly, as shown by Tarnow (1999), some occurrences of honorary authors can be caused by the will to maintain social relations. In a recent study, ghost authorship was also discussed in relation to accountability; especially in industrial biomedical research. For instance, Gøtzsche et al. (2007) found evidence of ghost authors in about 33 (75%) of the 44 industry-
initiated clinical trials they studied. In these studies, industries should have signed the paper but did not. Similar to ghost authors, Kwok (2005) defined a new type of colleague, the *White Bull*, who uses his seniority to distort authorship credit, generally at the expense of junior colleagues. Given their low status in the hierarchy of science, students are not likely to be honorary or guest authors. And the hypothesis that they can be sidelined because of ghost authorship if they work in industries seems quite remote, although difficult to assess. On the other hand, doctoral students are more likely to suffer from working with *white bull* colleagues (Kwok, 2005), given their weak negotiating power.

In a qualitative study of French researchers, Pontille (2004) revealed the differences between authorship practices in the fields of law, biology and sociology. In law, given that the research activity is rarely empirical and generally structured around the interpretation of texts, collaborative research is the exception rather than the norm. Hence, papers are generally signed by single authors, who can be considered as being responsible for the content of the paper as a whole. Given the *archetype* of the authorial figure, authors are also less disposed to grant co-authorship to subordinates—even when they contributed to the research. Hence, in this highly hierarchical field, the order in which authors are presented is generally a reflection of individuals’ status rather than of their input. The only exception to this is when graduate students have to publish a paper in order to fulfill the requirements for a position.
In biology, there are two positions that matter (Pontille, 2004): the first and the last author. The first author is the individual who did most of the experimental work—generally the graduate student—and the last author is the director of the laboratory, who gets to sign each and every paper done in his or her lab. In the middle of these two signatures will appear the names of technicians, other graduate students or researchers who contributed to technical or methodological aspects of the research. If they contributed to the analysis or design of the experiments, they will move closer to the first author if they are junior researchers or closer to the last author if they are senior researchers. So, in essence, biologists grant authorship to a broad spectrum of individuals. However, those who get credit for a discovery are normally the first(s) and last(s) authors.

Given the cohabitation of theoretical and empirical research in sociology, two general trends emerge (Pontille, 2004). In theoretical sociology, authorship practices are quite similar to those in law. The, generally single, author is the individual who wrote the text; other contributions to the research will be stated in the acknowledgements. In empirical sociology, where research is generally organized in groups, involvement in data collection and analysis can grant authorship. In these cases, the first author is usually the individual responsible for the research project and the writing; the order of other contributors depends on their negotiating power.

In sum, these studies illustrate that authorship attribution is a complex phenomenon that varies considerably among disciplines. We can, however, conclude that, in general, authorship is a measure of an individual’s integration and involvement in
research. The extent of this integration varies greatly by discipline, given the differences in authorship attribution practices. In the SSH, it is very likely that article-based bibliometric data of this thesis measures the ‘tip of the iceberg’, given that, as shown by Pontille (2004), only the writing of a paper grants authorship. There will likely be significant differences between SSH fields, given that quantitative-fields are more likely to provide collaboration opportunities than qualitative fields (Moody, 2004) and that, as a consequence, PhD students are more likely to collaborate with their advisors in the former than in the latter. Conversely, in the NSE, bibliometric data could overestimate the phenomenon, since involvement as a technical assistant often grants authorship. Hence, data obtained for the NSE will not necessarily be comparable with that of the SSH.

2.4. Bibliometrics

The term bibliometrics was first coined by Alan Pritchard in 1969 (Pritchard, 1969). It can be generally defined as “a field that uses mathematical and statistical techniques, from counting to calculus, to study publishing and communication patterns in the distribution of information” (Diodato, 1994). In other words, bibliometrics is the “…application of various statistical analyses to study patterns of authorship, publication, and literature use” (Lancaster, 1977). Though these definitions could apply to the study of any kind of literature, from novels to newspapers, bibliometrics is generally used to measure science and technology (Moed, 2005; van Raan, 1988). It uses published scientific literature (articles, books, conference proceedings, etc.) as a way of measuring scientific activity.
Until the 1960s, bibliometrics was mainly used by librarians to help them cope with the increasingly large number of journals published, both in terms of shelf space and ‘size’ of scientific information (Archambault and Larivière, 2007). Researchers such as Gross and Gross (1927), Alfred J. Lotka (1926) and Samuel Clement Bradford (1934) made seminal discoveries using small datasets compiled by hand. This practice changed in the 1960s with the arrival of computers and databases. Indeed, the modern history of bibliometrics is strongly linked with that of a scientific bibliographic database—the Science Citation Index (SCI)—created in 1963 by Eugene Garfield (Wouters, 1999). This bibliographic database has two characteristics that make it the most commonly used database for bibliometrics. First, it indexes references made by each source item it files, which allows for the compilation of citation counts and, hence, scientific impact measures (Garfield, 1979) and co-citation networks (Small, 1973). Second, it indexes all institutional addresses of authors, which makes it possible to regionalize research activity. In addition to who, these addresses allow us to answer where. Addresses also make it possible to compile data on international and interinstitutional collaboration. This is not the case with other bibliographic databases such as Medline, which only files the first address appearing on the paper. One of the postulates of bibliometrics is that the addresses appearing on papers are a good indicator of the place where the research was conducted. Thus, three aspects of scientific production can be measured with bibliometrics: 1) the institutional context of the production (the research institution(s) where new knowledge is produced), 2) the individual status of authors (faculty, students, postdoctoral fellows, etc.) and 3) the fields and subfields of research.
Though bibliometrics’ origins and first applications were in library science, recent works in the field have been produced by both information scientists and sociologists of science (Glänzel, 2003). Hence, the methodological roots of bibliometric research are in library and information sciences—since it uses databases of scientific documents as its main source of data—but as bibliometrics is generally applied to scientific documents, its results are often of great interest to sociologists of science. Bibliometrics can, nonetheless, now be considered a field in itself—the science of science. Indeed, since 1978, scientometricians have their own journal, *Scientometrics* and, since 1993, their own international scientific society, the International Society for Scientometrics and Informetrics (ISSI)\(^\text{10}\). The foundations of both a journal and a scientific society are, according to Gingras (1991), two criteria which denote that a field has attained the status of an autonomous discipline or specialty.

One of the basic ideas of bibliometrics is that new knowledge created by scientists is embedded in the scientific literature, and that by measuring scientific literature, we can measure knowledge. In fact, it is well known that scientists do not keep their discoveries for themselves. Scientific discoveries are transmitted in written form (an article or a book), and their *validity*, defined as their acceptance by the scientific community as a valid body of knowledge, is assured by peer review (Merton, 1973). In that sense, as shown in the preceding section, the signature of a research paper by a student or a researcher can be a measure of that individual’s contribution to the advancement of scientific knowledge, given the fact that signing a paper attributes to

\(^{10}\) [www.issi-society.info](http://www.issi-society.info)
the author the credit for the research results. However, the bibliometric methods used might minimize the contribution of students, since students depend on their advisors in so many aspects and might not negotiate authorship attribution as equals. Hence, they might not always appear as authors on the papers to which they contribute (Costa and Gatz, 1992; Sandler and Russell, 2005). For example, graduate students could participate in the compilation of data or in an experiment, yet not sign the resulting paper. In that sense, the participation in peer-reviewed papers could be considered as an indicator of doctoral students’ formal contribution to research (Borgman, 1990).

2.4.1. Limitations of bibliometrics and differences among disciplines

There are also several limits to bibliometric methods, which are a function of differences in research practices and publishing patterns among SSH and NSE. For this thesis, the main limitation is that current bibliometric tools do not count each and every publication in which doctoral students might participate. This limitation is especially important in the social sciences, and even more in the humanities. More specifically, the application of article counting methods to the social sciences and the humanities poses two main problems: 1) existing large-scale databases such as WOS or Scopus only index journal articles and, hence, exclude other type of documents—such as books and book chapters—and 2) existing databases have limited coverage of articles written in languages other than English. That being said, bibliometric databases focus on the scientific contributions that were published in core journals of

---

11 This section of the chapter presents overall limitations of bibliometrics. Limitations of specific indicators, such as the impact factor, are presented in section 3.5.3.
their respective fields (Garfield, 1990), that is, the journals that are most cited in other journals.

The first limitation is determined by discipline-specific dissemination media: although bibliometric databases only index journal articles, scholars in the different disciplines do not make the same use of scientific journals to disseminate their research results. Indeed, several authors have analyzed disciplinary differences in the use of journal articles for research dissemination. Small and Crane (1979) found that 39% and 24.5% of items cited by journals in sociology and economics, respectively, were books. At the opposite side of the spectrum, only 0.9% of all references were to books in high-energy physics. Similarly, Leydesdorff (2003) found that 79% of references by articles indexed in the Science Citation Index (SCI) were made to other papers indexed in the database, but that this percentage was only 45% in the Social Science Citation Index (SSCI). Glänzel and Schoepfli (1999) also had similar results with Thomson Reuters’ databases; the percentage of references to journals they obtained varied from 94% in immunology to 35% in history and philosophy of science and social sciences.

Providing data for the entire spectrum of disciplines and for a long time period (1980-2000), Lariviere et al. (2006) conducted a similar survey, which showed that, in 2000, journals accounted for about 93% of the cited references in medical disciplines, and to a percentage between 68% and 87% in the various disciplines of the natural sciences. On the other hand, the percentage of references made to journals was about 50% in the social sciences and humanities. Among the social sciences,
psychology, law and economics were the disciplines with the highest proportion of references to serials, respectively 57%, 58% and 70% in 2000. In the humanities, however, this percentage was lower (33% in history, 27% in other humanities and 21% in literature). As a consequence of this limitation, the participation of doctoral students in peer-reviewed articles is likely to be strongly tied to the discipline in which they are active.

The second limitation is attributable to SSH scholars’ preferred language of knowledge dissemination as well as their research objects (Gingras, 1984). Although research objects in the natural and medical sciences are more ‘universal’—the electron behaves in the same manner in Canada and in Brazil for instance—the social sciences and the humanities are more local (families in South America, information-seeking behaviour of second year engineering students, etc.). These different research objects have an effect on the way research is conducted, but also on its diffusion. As a result, the audience of research in the social sciences and humanities is often more limited to a specific country or region in what could be called ‘national’ journals (Glänzel, 1996; Hicks, 1999; Hicks, 2004; Nederhof et al., 1989; Webster, 1998; Winelawska, 1996). In other words, the ‘objects’ studied by the SSH are often more local and, hence, SSH researchers generally publish in their mother language and in local journals (Gingras, 1984; Line, 1999).

In order to measure national differences in the coverage of journals by Thomson Scientific, Archambault et al. (2006) compared journals indexed by Thomson Reuters’ databases to the exhaustive list of peer-reviewed journals from all over the world
(Ulrich’s periodicals directory\textsuperscript{12}). They found that, in the social sciences and the humanities (SSCI and AHCI databases), editors of journals indexed in Thomson’s databases were 55% more likely to be from the United Kingdom than in Ulrich and 35% more likely to be from the United States. On the other hand, journal editors from Canada, France and Germany were under-represented.

Similarly, comparison of the language of journals indexed by Thomson with the language identified by Ulrich revealed a bias in favour of English-language journals. Indeed, while 75% of peer-reviewed journals indexed in Ulrich are in English, this percentage is 90% for those indexed in Thomson Reuters’ databases. This evidence showed that for the combined SSCI and AHCI coverage, there is a 20-25% bias in favour of English-language scientific output in the SSH, and that French, German, and Spanish journals are underestimated by 28%, 50%, and 69%, respectively.

Although Canada was not significantly underestimated, we can expect Quebec, because of French language journals, to be more underestimated than the rest of Canada. That being said, this underestimation should not affect our results, since only papers from Quebec are analyzed in this thesis. Moreover, given that the database contains the core journals and that one could expect that students’ names could appear more often as collaborators in local journals, we can conclude that the results obtained provide a lower bound.

In addition to these well-known limitations in terms of coverage, other bibliometric differences can also be observed between disciplines. In addition to differences in

\textsuperscript{12} http://www.ulrichsweb.com
average productivity measures (see figures 14 and 38), the pace of research is also quite different in the various disciplines, which has a strong influence on the various bibliometric indicators. For instance, in the natural sciences and medical sciences, the pace at which these citations are received is much faster than in the social sciences and the humanities. As shown by Glänzel and Schoepflin (1999), the mean number of references per paper varies greatly among disciplines—from 14.9 in information and library science to 48.7 in history and philosophy of science and social science—and so does the average age of cited references. Indeed, while the average age of cited documents was 6.9 and 7.9 years in immunology and research medicine respectively, it was 12.5 years in sociology and close to 40 years in history and philosophy of science and social science.

Along the same lines, the average number of citations subsequently received by papers varies greatly by discipline, which is why citations measures have to take into account the disciplines of the journals in which the papers are published (Moed, De Bruin and van Leeuwen, 1995; Schubert and Braun, 1986). As shown by Wallace, Larivière and Gingras (2009), the average number of citations received by papers published in 2004 two years after publication was 6.3 in medicine, 2.8 in natural sciences and engineering and 1.9 in the social sciences. Similar differences are also observed in terms of ‘uncitedness’, that is, the percentage of papers that are not yet cited. In medicine, 20% of papers are not yet cited 10 years after their publication, while this percentage is 25% and 30% in natural sciences and engineering and in the social sciences. In the humanities, this percentage is much higher, with more than 80% of papers not yet cited 5 years after publication (Larivière et al., 2008). The main
implication of these different citation rates is that it is necessary to use a longer citation window when examining the impact of research in the social sciences or the humanities.

Finally, another important difference between disciplines which can be measured bibliometrically is research collaboration. In an analysis of collaboration practices among disciplines, Larivière, Gingras and Archambault (2006) showed that, in 2004, more than 90% of Canadian papers were co-authored in the natural sciences. This percentage was 68% in the social sciences, but only slightly above 10% in the arts and the humanities. Analysis of interinstitutional and international collaborations also revealed similar patterns. Along the same lines, although in both natural and social sciences a significant increase in the various types of collaboration was observed between 1980 and 2004, no such increase was observed in the humanities.

In sum, these bibliometric differences across disciplines influence the manner in which the various indicators are compiled and normalized, as well as provide some clues regarding the disciplinary differences observed in terms of students’ participation in papers. The various indicators used in this thesis are presented in section 3.6.
3. Research methods

As one could expect from reading previous studies on students’ contribution to papers, the methodological challenges for compiling such bibliometric data at the macro level are numerous. This section outlines the research method used in this doctoral research. It first presents the bibliometric database used as well as the database of students provided by the Ministère de l’éducation, du loisir et du sport du Québec (MELS), namely the Gestion des données sur l’effectif universitaire (GDEU). It then discusses some of the previous attempts at compiling bibliometric data at the level of individual researchers and describes the methods used for isolating scientific papers authored by PhD students (algorithm as well as manual validation). Finally, the bibliometric indicators and statistical tests used in this thesis are presented.

The main dataset analyzed in this thesis is obtained by merging two heterogeneous sources of information: the database of all published scientific papers in Quebec and the database of all graduate students in Quebec using the names of authors and of students (Figure 1). As presented in the last section, databases used for bibliometric analysis were not designed for that purpose, but for bibliographical research. Thus, they have two shortcomings when it comes to compiling statistics on individual authors. First, bibliometric databases do not include the first names of authors of
papers but only their initials\textsuperscript{13}. In other words, in the bibliographic record of his papers, John William Dawson\textsuperscript{14}—who is obviously absent from the students’ lists as well as from the Quebec papers database, would appear as Dawson-JW or Dawson-J (Figure 2)\textsuperscript{15}, which creates many homographic problems, especially for very common surnames (Figure 3). Hence, each paper retrieved for a given researcher has to be analyzed in order to remove the papers written by homographs. The same applies for the citations John William Dawson might receive. The other limitation is caused by the fact that the database does not contain any information on the relationship between the authors’ names (Allen-DS, Smith,-J, etc.) and their institutional addresses (McGill University, Harvard University, etc.). Thus, for a given paper signed by John, Jack and Paul and on which McGill, Harvard and Toronto universities appear, it is impossible to know which researcher belongs to which research organization. The absence of such links causes several problems, especially in the context of papers written in collaboration—and this is generally the case (Larivière, 2007)—, as a researcher from McGill University might collaborate with a colleague from another university who has the same name and initials as another McGill researcher (homography of collaborators). Hence, isolating publications from

\textsuperscript{13} Although Thomson Reuters now index the full first name and initials of the authors and provide a link between each of the authors and its institution of affiliation, this was not available at the time the data for this thesis were compiled. Thomson Reuters does not yet provide this data retrospectively. As recently discussed by Enserink (2009), both Thomson Reuters and Scopus are currently trying to implement a unique researcher ID, which would greatly help finding papers from individual authors. However, there are several competing initiatives for the moment, which hinders the development of this no-so-unique researcher ID.

\textsuperscript{14} Sir John William Dawson was McGill’s principal from 1855 to 1893. He is generally considered as the “father” of modern McGill.

\textsuperscript{15} Along the same lines, several journals only list the initials of the authors because of constraints of space—especially in physics where author lists are very long.
doctoral students using bibliometric databases at the macro level is a complex task, which explains the lack of such a study so far.

**Figure 1. Creation of the database of scientific papers to which doctoral students contributed**
3.1. Bibliometric database

The database of scientific papers that is used in this thesis has been constructed with the Web of Science version of the Science Citation Index Expanded (SCIE), the Social...
Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI)—grouped under the Web of Science—of the Institute for Scientific Information (now Thomson Reuters). These three databases index articles published in about 9,000 international peer-reviewed journals and cover the 1980–2007 period for a total of about 1 million document per year. Thomson Reuters’ bibliographic databases are primarily made for bibliographic searches. Hence, the online version of these databases has several limitations for bibliometric analyses, as the user can only compile statistics using a few preselected variables. For example, it is impossible, using the web interface of the WOS, to compile data on the scientific production of all authors of each of Quebec’s universities, or to analyze results based on datasets larger than 100,000 records.

In order to compile advanced bibliometric indicators such as the ones presented in this thesis, one has to extract all the source data from the WOS and transform them into a relational database. The version of the WOS used in this thesis has been built by the Observatoire des sciences et des technologies (OST) using the source data in XML format provided by Thomson Reuters. The XML source data was transformed into a relational database stored on a Microsoft SQL-Server (see Figure 4), with which any variable can be linked to any other.
Since the OST created this database in 1997, many improvements have been made to the database, among which the creation of an authority table for all Canadian research institutions is probably the most significant. Indeed, Thomson Reuters’s databases do not have a ‘preferred form of heading’ for research institutions, which means that publications from McGill University could be labelled as MacGill-Univ, Mcill-Univ, etc. OST’s team went through all Canadian organizations from all sectors—universities, hospitals, governments and industries (N=15,000)—in order to
assign them a preferred form of heading. Using the authority table, it is thus possible in this thesis to analyse the collaboration patterns of Quebec’s students with governments, industries, etc.

This database does not cover all published research. Some scientific discoveries can be diffused in non-indexed journals, but also in other types of documents such as conference proceedings, grey literature and books. This database nonetheless contains research that is published in international core journals, i.e., journals that are most visible to both national and international scientific communities and that have the highest scientific impact in their respective subfields (Garfield, 1990). Though there is some controversy on this selection criteria and on the English-language bias of the original journal dataset (Archambault and Larivière, 2007), Thomson Reuters’ can still be considered as the staple database for bibliometric analysis—although the Scopus database, created by Elsevier in 200416, is challenging this status17. Indeed, the Scopus database has basically the same features as the WOS, although it does not go back as far in time for previous years. The WOS has existed continuously since 1963, while Scopus can only be considered as ‘complete’ starting in 1996. Along the same lines, the WOS covers all disciplines, while Scopus started to index journals in the humanities in 2009.

---

16 [http://info.scopus.com/overview/what](http://info.scopus.com/overview/what)
17 A quick search for bibliometric papers performed on October 10th 2009 showed that, since 2005, the number of papers written in the journal *Scientometrics* using the Web of Science—or one of its databases—was more than 10 times higher than the number of papers using Scopus (116 vs. 11).
The selection criteria of each database are also different. While Scopus indexes all Elsevier journals as well as any other journal that might want to be indexed, the WOS has a standard journal selection process, in which both quantitative (such as citations and diversity of countries of authors) and qualitative characteristics are taken into account\(^\text{18}\). Hence, very local or low-tier journals are generally excluded from the WOS, but might be indexed in Scopus if they send their bibliographic information to Elsevier. Because of its status as the staple database, its robustness and well-known characteristics (both positive and negative) as well as for reasons of comparability with most bibliometric studies made worldwide, it was decided to use the WOS. That being said, the overlap between the two databases is very high—but as one might expect, varies between fields—and the correlation between publication and citation measures obtained at the level of countries and disciplines is high, with \(R^2\) generally greater than 0.99 (Archambault \textit{et al.}, 2009).

Thomson Reuters’ databases list several types of documents (articles, book reviews, editorials). However, we do not consider them all to be contributions to the advancement of science. In most bibliometric studies, as in this thesis, only articles, research notes and review articles are used because they are the main modes for diffusing new knowledge (Moed, 1996). However, there are no clear standards on this subject and, in some disciplines, other types of documents are deemed to make a contribution to the advancement of knowledge, while in others they are not. For example, meeting abstracts and conference proceedings are often considered original research articles in the fields of engineering and computer science (Godin, 1998; 18 http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process/
Visser and Moed, 2005; Lisée, Larivière and Archambault, 2008). As a consequence, the bibliometric data analyzed in this thesis might not be as precise in engineering as it is for physics or for medical disciplines. The same statement also applies to the social sciences and humanities. Nevertheless, in those fields, the bibliometric data is still representative of the subset of doctoral researchers who publish articles.

3.1.1 Disciplinary classification of journals

In order to have a disciplinary breakdown of publications assigned to doctoral students, as well as to help both the algorithm-based and manual validation of doctoral students’ scientific papers, each journal of the database was categorized into a discipline and a specialty according to the classification scheme used by U.S. National Science Foundation (NSF) in its Science and Engineering Indicators series\(^{19}\). The main advantage of this classification scheme over that provided by Thomson Reuters is that 1) it has a two-level classification (discipline and specialty), which allows the use of two different levels of aggregation and, 2) it categorizes each journal into only one discipline and specialty, which prevents double counts of papers when they are assigned to more than one discipline.

The NSF classification scheme, however, does not provide any disciplinary classification for journals in the arts and humanities—mainly journals indexed by the AHCI. Hence, these journals were manually categorized by OST in specialties of the arts and the humanities. Tables 3 and 4 present the list of disciplines and specialties as well as the number of journals and papers they include for SSH and NSE,

\(^{19}\) More details on the classification scheme can be found at: http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1
respectively. In the SSH (Table 3), the classification groups 45 specialties into 6 disciplines. This number is higher in the NSE: 98 specialties are grouped into 8 disciplines.

As one might expect from comparing the publication patterns of social scientists to those of natural scientists, the number of journals and papers categorized in SSH is quite a bit smaller than it is in the NSE. Indeed, while close to 686,000 papers are published in about 3,500 SSH journals over the 2000-2007 period, these numbers are, in NSE, almost 10 times higher for papers (6,417,08) and two times higher for journals (7,649). The average number of papers per journal is also quite different: while about 200 papers per journal were published in SSH during the period, this number is more than four times higher in NSE (≈840). In NSE, medicine-related disciplines (biomedical research and, especially, clinical medicine) have a lot of importance and account for 44% of all papers and journals (33% of journals and 31% of papers for clinical medicine alone). In SSH, however, the number of papers and journals is more evenly dispersed across disciplines. Finally, four journals could not be categorized into a discipline and specialty, for a total of 2,215 articles. The papers published in these journals were thus excluded from the analysis. One can also note that some of the specialties are redundant—such as social sciences, biomedical and social studies of medicine, which would benefit from being merged, as both categories contain a small number of journals relating to the social aspects of health and medicine.²⁰

²⁰ The disciplinarity classification of journals was updated in the 2010 Science and engineering indicators report (NSF, 2010).
Table 2. Disciplinary classification of SSH journals (number of journals and papers presented in parenthesis), 2000-2007

<table>
<thead>
<tr>
<th>Disciplinary Field</th>
<th>Journals</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>**Arts (200</td>
<td>17,108)**</td>
<td></td>
</tr>
<tr>
<td>Fine Arts &amp; Architecture (98</td>
<td>9,628)</td>
<td></td>
</tr>
<tr>
<td>Performing Arts (102</td>
<td>7,480)</td>
<td></td>
</tr>
<tr>
<td>**Health (348</td>
<td>102,754)**</td>
<td></td>
</tr>
<tr>
<td>Geriatrics &amp; Gerontology (35</td>
<td>8,854)</td>
<td></td>
</tr>
<tr>
<td>Health Policy &amp; Services (72</td>
<td>20,368)</td>
<td></td>
</tr>
<tr>
<td>Nursing (70</td>
<td>17,035)</td>
<td></td>
</tr>
<tr>
<td>Public Health (82</td>
<td>30,066)</td>
<td></td>
</tr>
<tr>
<td>Rehabilitation (56</td>
<td>14,537)</td>
<td></td>
</tr>
<tr>
<td>Social Sciences, Biomedical (11</td>
<td>5,861)</td>
<td></td>
</tr>
<tr>
<td>Social Studies of Medicine (2</td>
<td>715)</td>
<td></td>
</tr>
<tr>
<td>Speech-Language Pathol, and Audiol. (20</td>
<td>5,318)</td>
<td></td>
</tr>
<tr>
<td>**Humanities (892</td>
<td>106,810)**</td>
<td></td>
</tr>
<tr>
<td>History (224</td>
<td>25,805)</td>
<td></td>
</tr>
<tr>
<td>Language &amp; Linguistics (140</td>
<td>16,650)</td>
<td></td>
</tr>
<tr>
<td>Literature (267</td>
<td>30,165)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Humanities (84</td>
<td>9,834)</td>
<td></td>
</tr>
<tr>
<td>Philosophy (98</td>
<td>14,886)</td>
<td></td>
</tr>
<tr>
<td>Religion (79</td>
<td>9,470)</td>
<td></td>
</tr>
<tr>
<td>**Professional Fields (663</td>
<td>136,239)**</td>
<td></td>
</tr>
<tr>
<td>Communication (33</td>
<td>5,545)</td>
<td></td>
</tr>
<tr>
<td>Education (166</td>
<td>29,132)</td>
<td></td>
</tr>
<tr>
<td>Information Science &amp; Library Sci. (73</td>
<td>13,244)</td>
<td></td>
</tr>
<tr>
<td>Law (109</td>
<td>17,249)</td>
<td></td>
</tr>
<tr>
<td>Management (223</td>
<td>58,946)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Professional Field (27</td>
<td>5,180)</td>
<td></td>
</tr>
<tr>
<td>Social Work (32</td>
<td>6,943)</td>
<td></td>
</tr>
<tr>
<td>**Psychology (437</td>
<td>132,123)**</td>
<td></td>
</tr>
<tr>
<td>Behavioral Science &amp; Complementary Psycho. (32</td>
<td>18,596)</td>
<td></td>
</tr>
<tr>
<td>Clinical Psychology (62</td>
<td>17,922)</td>
<td></td>
</tr>
<tr>
<td>Developmental &amp; Child Psychology (55</td>
<td>15,728)</td>
<td></td>
</tr>
<tr>
<td>Experimental Psychology (58</td>
<td>19,800)</td>
<td></td>
</tr>
<tr>
<td>General Psychology (51</td>
<td>11,223)</td>
<td></td>
</tr>
<tr>
<td>Human Factors (16</td>
<td>6,136)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Psychology (90</td>
<td>24,229)</td>
<td></td>
</tr>
<tr>
<td>Psychoanalysis (17</td>
<td>3,816)</td>
<td></td>
</tr>
<tr>
<td>Social Psychology (56</td>
<td>14,673)</td>
<td></td>
</tr>
<tr>
<td>**Social Sciences (956</td>
<td>191,661)**</td>
<td></td>
</tr>
<tr>
<td>Anthropology and Archaeology (111</td>
<td>15,865)</td>
<td></td>
</tr>
<tr>
<td>Area Studies (70</td>
<td>10,379)</td>
<td></td>
</tr>
<tr>
<td>Criminology (31</td>
<td>5,526)</td>
<td></td>
</tr>
<tr>
<td>Demography (21</td>
<td>3,431)</td>
<td></td>
</tr>
<tr>
<td>Economics (214</td>
<td>59,243)</td>
<td></td>
</tr>
<tr>
<td>General Social Sciences (62</td>
<td>12,664)</td>
<td></td>
</tr>
<tr>
<td>Geography (68</td>
<td>13,744)</td>
<td></td>
</tr>
<tr>
<td>International Relations (43</td>
<td>7,459)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Social Sciences (79</td>
<td>15,100)</td>
<td></td>
</tr>
<tr>
<td>Planning &amp; Urban Studies (34</td>
<td>6,122)</td>
<td></td>
</tr>
<tr>
<td>Political Science and Public Administration (94</td>
<td>17,615)</td>
<td></td>
</tr>
<tr>
<td>Science studies (33</td>
<td>6,155)</td>
<td></td>
</tr>
<tr>
<td>Sociology (96</td>
<td>18,358)</td>
<td></td>
</tr>
<tr>
<td>**All SSH Disciplines (3,496</td>
<td>686,695)**</td>
<td></td>
</tr>
</tbody>
</table>

57
Table 3. Disciplinary classification of NSE journals (number of journals and papers presented in parenthesis), 2000-2007

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Number of Journals</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology (943</td>
<td>508606)</td>
<td></td>
</tr>
<tr>
<td>Agricut &amp; Food Science (233</td>
<td>127416)</td>
<td></td>
</tr>
<tr>
<td>Botany (227</td>
<td>123500)</td>
<td></td>
</tr>
<tr>
<td>Dairy &amp; Animal Science (42</td>
<td>28559)</td>
<td></td>
</tr>
<tr>
<td>Ecology (80</td>
<td>49013)</td>
<td></td>
</tr>
<tr>
<td>Entomology (74</td>
<td>33922)</td>
<td></td>
</tr>
<tr>
<td>General Biology (31</td>
<td>18382)</td>
<td></td>
</tr>
<tr>
<td>General Zoology (91</td>
<td>31368)</td>
<td></td>
</tr>
<tr>
<td>Marine Biology &amp; Hydrobiology (94</td>
<td>62520)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Biology (35</td>
<td>20270)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Zoology (36</td>
<td>13656)</td>
<td></td>
</tr>
<tr>
<td>Biomedical Research (866</td>
<td>850235)</td>
<td></td>
</tr>
<tr>
<td>Anatomy &amp; Morphology (11</td>
<td>5365)</td>
<td></td>
</tr>
<tr>
<td>Biomedical Engineering (47</td>
<td>38722)</td>
<td></td>
</tr>
<tr>
<td>Biophysics (22</td>
<td>16144)</td>
<td></td>
</tr>
<tr>
<td>Cellular Biology Cyto &amp; Histology (80</td>
<td>58625)</td>
<td></td>
</tr>
<tr>
<td>Embryology (11</td>
<td>10488)</td>
<td></td>
</tr>
<tr>
<td>General Biomedical Research (89</td>
<td>99779)</td>
<td></td>
</tr>
<tr>
<td>Genetics &amp; Heredity (114</td>
<td>92397)</td>
<td></td>
</tr>
<tr>
<td>Microbiology (101</td>
<td>87043)</td>
<td></td>
</tr>
<tr>
<td>Microscopy (5</td>
<td>4357)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Biomedical Research (24</td>
<td>20093)</td>
<td></td>
</tr>
<tr>
<td>Nutrition &amp; Dietetic (54</td>
<td>31068)</td>
<td></td>
</tr>
<tr>
<td>Parasitology (21</td>
<td>14705)</td>
<td></td>
</tr>
<tr>
<td>Physiology (45</td>
<td>49595)</td>
<td></td>
</tr>
<tr>
<td>Virology (23</td>
<td>30607)</td>
<td></td>
</tr>
<tr>
<td>Chemistry (491</td>
<td>776515)</td>
<td></td>
</tr>
<tr>
<td>Analytical Chemistry (68</td>
<td>87357)</td>
<td></td>
</tr>
<tr>
<td>Applied Chemistry (22</td>
<td>16878)</td>
<td></td>
</tr>
<tr>
<td>General Chemistry (120</td>
<td>155729)</td>
<td></td>
</tr>
<tr>
<td>Inorganic &amp; Nuclear Chemistry (31</td>
<td>55654)</td>
<td></td>
</tr>
<tr>
<td>Organic Chemistry (52</td>
<td>128718)</td>
<td></td>
</tr>
<tr>
<td>Physical Chemistry (112</td>
<td>242934)</td>
<td></td>
</tr>
<tr>
<td>Polymers (86</td>
<td>89245)</td>
<td></td>
</tr>
<tr>
<td>Clinical Medicine (2529</td>
<td>1973254)</td>
<td></td>
</tr>
<tr>
<td>Addictive Diseases (29</td>
<td>14696)</td>
<td></td>
</tr>
<tr>
<td>Allergy (10</td>
<td>9800)</td>
<td></td>
</tr>
<tr>
<td>Anesthesiology (22</td>
<td>22971)</td>
<td></td>
</tr>
<tr>
<td>Arthritis &amp; Rheumatology (27</td>
<td>39290)</td>
<td></td>
</tr>
<tr>
<td>Cancer (126</td>
<td>134335)</td>
<td></td>
</tr>
<tr>
<td>Cardiovascular System (122</td>
<td>98749)</td>
<td></td>
</tr>
<tr>
<td>Dentistry (60</td>
<td>36976)</td>
<td></td>
</tr>
<tr>
<td>Dermatology &amp; Veneral Disease (48</td>
<td>36195)</td>
<td></td>
</tr>
<tr>
<td>Endocrinology (74</td>
<td>66377)</td>
<td></td>
</tr>
<tr>
<td>Environmental &amp; Occupational Health (49</td>
<td>29199)</td>
<td></td>
</tr>
<tr>
<td>Fertility (16</td>
<td>16686)</td>
<td></td>
</tr>
<tr>
<td>Gastroenterology (55</td>
<td>55011)</td>
<td></td>
</tr>
<tr>
<td>General &amp; Internal Medicine (235</td>
<td>153995)</td>
<td></td>
</tr>
<tr>
<td>Geriatrics (13</td>
<td>9614)</td>
<td></td>
</tr>
<tr>
<td>Hematology (51</td>
<td>45886)</td>
<td></td>
</tr>
<tr>
<td>Immunology (130</td>
<td>133770)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Clinical Medicine (76</td>
<td>29303)</td>
<td></td>
</tr>
<tr>
<td>Nephrology (18</td>
<td>23216)</td>
<td></td>
</tr>
<tr>
<td>Neurology &amp; Neurosurgery (243</td>
<td>229728)</td>
<td></td>
</tr>
<tr>
<td>Obstetrics &amp; Gynecology (61</td>
<td>38219)</td>
<td></td>
</tr>
<tr>
<td>Ophthalmology (43</td>
<td>41774)</td>
<td></td>
</tr>
<tr>
<td>Orthopedies (47</td>
<td>42405)</td>
<td></td>
</tr>
<tr>
<td>Otorhinolaryngology (37</td>
<td>28816)</td>
<td></td>
</tr>
<tr>
<td>Pathology (42</td>
<td>31629)</td>
<td></td>
</tr>
<tr>
<td>Pediatrics (58</td>
<td>41337)</td>
<td></td>
</tr>
<tr>
<td>Pharmacology (239</td>
<td>169647)</td>
<td></td>
</tr>
<tr>
<td>Pharmacy (16</td>
<td>15883)</td>
<td></td>
</tr>
<tr>
<td>Psychiatry (106</td>
<td>52719)</td>
<td></td>
</tr>
<tr>
<td>Radiology &amp; Nuclear Medicine (93</td>
<td>76973)</td>
<td></td>
</tr>
<tr>
<td>Respiratory System (31</td>
<td>27142)</td>
<td></td>
</tr>
<tr>
<td>Surgery (140</td>
<td>108364)</td>
<td></td>
</tr>
<tr>
<td>Tropical Medicine (14</td>
<td>9212)</td>
<td></td>
</tr>
<tr>
<td>Urology (49</td>
<td>39623)</td>
<td></td>
</tr>
<tr>
<td>Veterinary Medicine (149</td>
<td>78724)</td>
<td></td>
</tr>
<tr>
<td>Earth and Space (565</td>
<td>372435)</td>
<td></td>
</tr>
<tr>
<td>Astronomy &amp; Astrophysics (45</td>
<td>72999)</td>
<td></td>
</tr>
<tr>
<td>Earth &amp; planetary Science (117</td>
<td>81251)</td>
<td></td>
</tr>
<tr>
<td>Environmental Science (141</td>
<td>98775)</td>
<td></td>
</tr>
<tr>
<td>Geology (173</td>
<td>63872)</td>
<td></td>
</tr>
<tr>
<td>Meteorology &amp; Atmospheric Science (41</td>
<td>30231)</td>
<td></td>
</tr>
<tr>
<td>Oceanography &amp; Limnology (48</td>
<td>25307)</td>
<td></td>
</tr>
<tr>
<td>Engineering and Technology (1387</td>
<td>900689)</td>
<td></td>
</tr>
<tr>
<td>Aerospace Technology (24</td>
<td>17899)</td>
<td></td>
</tr>
<tr>
<td>Chemical Engineering (110</td>
<td>78147)</td>
<td></td>
</tr>
<tr>
<td>Civil Engineering (66</td>
<td>26376)</td>
<td></td>
</tr>
<tr>
<td>Computers (281</td>
<td>191363)</td>
<td></td>
</tr>
<tr>
<td>Electrical Engineering &amp; Electronics (263</td>
<td>169275)</td>
<td></td>
</tr>
<tr>
<td>General Engineering (18</td>
<td>4411)</td>
<td></td>
</tr>
<tr>
<td>Industrial Engineering (25</td>
<td>9811)</td>
<td></td>
</tr>
<tr>
<td>Materials Science (247</td>
<td>192681)</td>
<td></td>
</tr>
<tr>
<td>Mechanical Engineering (146</td>
<td>79194)</td>
<td></td>
</tr>
<tr>
<td>Metals &amp; Metallurgy (95</td>
<td>74908)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Engineering &amp; Technology (59</td>
<td>24584)</td>
<td></td>
</tr>
<tr>
<td>Nuclear Technology (23</td>
<td>24162)</td>
<td></td>
</tr>
<tr>
<td>Operations Research (30</td>
<td>7878)</td>
<td></td>
</tr>
<tr>
<td>Mathematics (461</td>
<td>210313)</td>
<td></td>
</tr>
<tr>
<td>Applied Mathematics (135</td>
<td>65314)</td>
<td></td>
</tr>
<tr>
<td>General Mathematics (221</td>
<td>102867)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Mathematics (24</td>
<td>11228)</td>
<td></td>
</tr>
<tr>
<td>Probability &amp; Statistics (81</td>
<td>30904)</td>
<td></td>
</tr>
<tr>
<td>Physics (407</td>
<td>825034)</td>
<td></td>
</tr>
<tr>
<td>Acoustics (19</td>
<td>19345)</td>
<td></td>
</tr>
<tr>
<td>Applied Physics (78</td>
<td>192857)</td>
<td></td>
</tr>
<tr>
<td>Chemical Physics (25</td>
<td>73710)</td>
<td></td>
</tr>
<tr>
<td>Fluids &amp; Plasmas (27</td>
<td>25063)</td>
<td></td>
</tr>
<tr>
<td>General Physics (123</td>
<td>253585)</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous Physics (11</td>
<td>5932)</td>
<td></td>
</tr>
<tr>
<td>Nuclear &amp; Particle Physics (30</td>
<td>88450)</td>
<td></td>
</tr>
<tr>
<td>Optics (66</td>
<td>82983)</td>
<td></td>
</tr>
<tr>
<td>Solid State Physics (28</td>
<td>83109)</td>
<td></td>
</tr>
<tr>
<td>All Disciplines (7649</td>
<td>6417081)</td>
<td></td>
</tr>
</tbody>
</table>
3.2. Doctoral students’ database

Given that the status of individual authors rarely appears in the byline of papers, or in bibliographical databases, a list of doctoral students is needed to find papers they authored. Although all Quebec universities have their own lists of enrolments, expected differences in format and content, as well as difficulties of access, discouraged their use. Hence, it was decided to use the Quebec’s government administrative database of university enrolments in Quebec, the Gestion des données sur l'effectif universitaire database (GDEU)\(^{21}\). This database contains the names—essential information for this thesis—of all university students in the province of Quebec. In order to find publications authored by doctoral students, the database of scientific papers written by authors from the province of Quebec and the database of students were matched using the name of the student. As mentioned at the beginning of this section, several of these papers are authored by homographs.

Access to the disaggregated data of the GDEU database is secured, given the personal nature of the information it contains. Hence, its access is generally restricted to MELS internal use. In order to have access to this database, one has to go through the Commission d’accès à l’information du Québec (CAI), which decides after review if access can be granted or not and with what conditions. The CAI has two criteria for deciding whether or not the Commission grants access to personal information. The researcher needs to prove that, first, the projected use is not trivial and that the goal of the research can be attained only if the personal information is communicated and second, that the personal information will be used in a manner which ensures its

confidential character. Given that this PhD thesis fulfilled these criteria and that having the full names of the PhD students was mandatory for this research, the CAI granted access to the subset of PhD students of the GDEU database for a period of one year, which allowed me to compile the data on doctoral students’ scientific publications.

However, as could be expected, the CAI decided that only part of the GDEU database could be provided by MELS. More specifically, the information on student’s gender, age and on province or country of origin was considered too confidential to be provided. Hence, this thesis does not provide any analysis by students’ age or status (domestic vs. foreign students). The case of gender is less problematic, since it can generally be known using given names. Finally, although the initial request asked for students’ data from 1990 to 2008, only data for 2000 onward could be provided because the names of PhD students were not included in the database prior to fall 2000.

Therefore, the dataset provided by MELS only contained PhD students that either 1) enrolled in a PhD program during the period between fall 2000 and winter 2008 or 2) enrolled before fall 2000 but were still registered or graduated at any point after 2000. Hence, students that completed their PhD in 1998 or 1999 are excluded from the dataset, as well as students who joined a PhD after winter 2008. On the whole, the list of doctoral students contains 27,397 distinct doctoral students, with 2,838 (10.4%) having changed program at least once during the period under study.
The following variables were included in the list of students:

- **Students' unique identifier**: A distinct numeric code with which each student can be identified.
- **Surname**: Surname of the student. Some students have compound names (Tremblay-Gagnon, etc.). All names were included in the database.
- **Given name**: Given names of the students. Again, in the case of compound names, all given names were included.
- **University enrolled**: One of the 17 universities in Quebec that have at least one PhD program\(^{22}\).
- **Program enrolled**: Name of the PhD program in which students enrolled.
- **First semester enrolled in the program**: Semester in which the student enrolled in the program (ex. fall 2004).
- **Last semester enrolled in the program**: Semester at which the student was last enrolled (ex. winter 2008).
- **Graduating year**: Year of graduation. This field is empty for students that have not yet graduated.
- **Graduating semester**: Semester of graduation (ex. winter 2008). The field is empty for students that have not yet graduated.
- **Graduating university**: University that granted the PhD. It is usually the same as the university enrolled, but not always, as universities sometimes have joint programs. This field is empty for students that have not yet graduated.

Table 4 presents my own record, as contained in the list of PhD students. One can see that no standardisation was performed by MELS neither in terms of program names nor of Faculty affiliation. The following section presents the treatments that were made on the PhD students’ list prior to the reconstitution of their publication files.

\(^{22}\) Out of the 18 universities in Quebec, only Bishop’s University does not have any PhD program.
Several additions and modifications were made to the list of doctoral studies prior to the match of their scientific articles in the bibliometric database. The first modification was to add a new ‘name’ field in which the name of all students was written into the form used by Thomson Reuters in the Author field of the bibliometric database. Hence, the name Larivière, Vincent was converted into LARIVIERE-V. In the case of composed given names (Jean-Pierre, Marie-Andrée, etc.), only the first initial was included in order to match the highest number of publications and, consequently, lower false negatives. For example, the name Robitaille, Jean-Pierre was converted to ROBITAILLE-J.

The second addition to the list was to categorize each doctoral student as either male or female. As this was performed for statistical purposes and not for the evaluation
of individual researchers, this work need not be 100% accurate or exhaustive. Hence, this was done using only the list of 8,742 distinct names given and matching it with several in-house as well as publicly available lists of names categorized by gender. In cases where the given name could either be given to a male or a female, it was attributed to the gender for which it is the most common. When no predominant gender could be found, the name was categorized as unknown.

The first match was done using the list of names of Quebec university professors and researchers categorized by gender\textsuperscript{23}. This list contained 3,433 distinct given names. This first automatic match categorized 1,418 doctoral students’ given names as either male or female, accounting for 16.2% of the names but 62.2% of the doctoral students (Table 5). The second automatic match was performed with a publicly available database: the database of names given to newborns in the province compiled by the Régie des rents du Québec (RRQ) since 2004\textsuperscript{24}. The list of names given in 2004 was used. It included 8,712 women’s names and 7,420 men’s names and assigned 1,092 doctoral students’ given names (12.5% of the names and 9.6% of the students). The third round of matching was done using several lists of given names by origin (French names, English names, etc.) found on Wikipedia\textsuperscript{25}. Taken together, these lists included 18,860 names, of which 764 were matched (8.7% of the names and 5.6% of the students). On the whole, these three rounds of automatic gender attribution provided a gender for only 37.5% of the names, but 77.3% of the

\textsuperscript{23} This list was provided by Quebec’s Ministère du développement économique, de l’innovation et de l’exportation (MDEIE) as well as the three funding councils: Fonds de la recherche en santé du Québec (FRSQ), Fonds québécois de la recherche sur la société et la culture (FQRSC) and Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

\textsuperscript{24} \url{http://www.rrq.gouv.qc.ca/en/enfants/Pages/banque_prenoms.aspx}

\textsuperscript{25} \url{http://fr.wikipedia.org/wiki/Liste_des_prenoms}
doctoral students, as these names were more common and more likely to be given to a high number of students. A final round of gender attribution was performed by searching the given name on the Internet using Google to find any indication on its gender. This long task provided a gender for 2,747 distinct names, which account for 31.4% of the names but only 12.5% of the doctoral students.

Table 5. Number and percentage of distinct given names and doctoral students to which a gender could be attributed, by method

<table>
<thead>
<tr>
<th>Method</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distinct given names</td>
<td>Distinct PhD students</td>
</tr>
<tr>
<td>List of professors</td>
<td>1 418</td>
<td>17 037</td>
</tr>
<tr>
<td>Google search</td>
<td>2 747</td>
<td>3 414</td>
</tr>
<tr>
<td>RRQ</td>
<td>1 092</td>
<td>2 624</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>764</td>
<td>1 527</td>
</tr>
<tr>
<td>No match</td>
<td>2 721</td>
<td>2 795</td>
</tr>
<tr>
<td>All methods</td>
<td>8 742</td>
<td>27 397</td>
</tr>
</tbody>
</table>

Table 6 presents the number and percentage of doctoral students by gender.

Comparing these numbers with the data for the full population of Quebec doctoral students by gender in 2003—the middle year of our students’ dataset—available in the statistical report of the Canadian Association for Graduate Studies (2008), shows that the gender distribution of the sample of doctoral students to which a gender could be attributed is quite similar to that of the population of doctoral students. This suggests that the automatic as well as the manual attribution of gender did not overestimate nor underestimate any of the genders.
A third modification performed on the students’ list was to regroup doctoral programs. Indeed, similar programs in different universities don’t generally have the same name. For example, the PhD program in psychology at UQAM is called Doctorat en Psychologie at UQAM as opposed to Ph.D. Psychologie at the Université de Montréal. Similarly, the same program in English-language universities will obviously be named in English. Even within a given university, the same program can be written in several different manners. McGill University, for instance, has 359 distinct doctoral programs in the list, but several of them are duplicates.

On the whole, 866 distinct programs were included in the students’ list. Hence, in order to 1) find regularities in the publication patterns of doctoral students from similar departments to help the reconstitution of doctoral students’ publication files and 2) ultimately breakdown the statistics obtained by the discipline of doctoral program, the doctoral programs were categorized into 42 specialties and 9 disciplines.

---

For example, the Ph.D. program in sociology is written in three different manners: PREP PH.D.,Sociology –T; PH.D.,Sociology –T; PH.D SOCIOLOGY.

---

Table 6. Number and percentage of doctoral students, by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>% excluding unknowns</th>
<th>% in CAGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>11 270</td>
<td>41,1%</td>
<td>45,8%</td>
</tr>
<tr>
<td>Male</td>
<td>13 332</td>
<td>48,7%</td>
<td>54,2%</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 795</td>
<td>10,2%</td>
<td>-</td>
</tr>
<tr>
<td>All students</td>
<td>27 397</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
This classification scheme used for categorizing doctoral programs was based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP) developed by the U.S. Department of Education's National Center for Education Statistics (NCES). This classification of programs was recently used in Larivière et al. (2009) for Quebec’s university-affiliated researchers. A CIP-based discipline could not be attributed to 232 doctoral students (0.8% of the dataset); these were excluded from the analysis. Table 7 presents the final classification scheme as well as the number of doctoral students in each of the disciplines and specialties.

Table 7. Classification of programs and number of doctoral students (in parenthesis), 2000-2007

<table>
<thead>
<tr>
<th>Program</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Medical Sciences (3,567)</strong></td>
<td></td>
</tr>
<tr>
<td>General Medicine (124)</td>
<td></td>
</tr>
<tr>
<td>Laboratory Medicine (2,410)</td>
<td></td>
</tr>
<tr>
<td>Medical Specialties (1,032)</td>
<td></td>
</tr>
<tr>
<td>Surgical Specialties (46)</td>
<td></td>
</tr>
<tr>
<td><strong>Business &amp; Management (1,389)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Education (1,536)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Engineering (4,502)</strong></td>
<td></td>
</tr>
<tr>
<td>Chemical Engineering (409)</td>
<td></td>
</tr>
<tr>
<td>Civil Engineering (574)</td>
<td></td>
</tr>
<tr>
<td>Electrical &amp; Computer Engineering (1,616)</td>
<td></td>
</tr>
<tr>
<td>Mechanical &amp; Industrial Engineering (813)</td>
<td></td>
</tr>
<tr>
<td>Other Engineering (1,197)</td>
<td></td>
</tr>
<tr>
<td><strong>Health Sciences (635)</strong></td>
<td></td>
</tr>
<tr>
<td>Dentistry (24)</td>
<td></td>
</tr>
<tr>
<td>Kinesiology / Physical Education (131)</td>
<td></td>
</tr>
<tr>
<td>Nursing (113)</td>
<td></td>
</tr>
<tr>
<td>Other Health Sciences (112)</td>
<td></td>
</tr>
<tr>
<td>Public Health &amp; Health Administration (215)</td>
<td></td>
</tr>
<tr>
<td>Rehabilitation Therapy (45)</td>
<td></td>
</tr>
<tr>
<td><strong>Humanities (3,919)</strong></td>
<td></td>
</tr>
<tr>
<td>Fine &amp; Performing Arts (728)</td>
<td></td>
</tr>
<tr>
<td>Foreign Languages Literature, Linguistics &amp; Area Studies (743)</td>
<td></td>
</tr>
<tr>
<td>French/English (890)</td>
<td></td>
</tr>
<tr>
<td>History (507)</td>
<td></td>
</tr>
<tr>
<td>Philosophy (516)</td>
<td></td>
</tr>
<tr>
<td>Religious Studies &amp; Vocations (603)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-Health Professional (1,569)</strong></td>
<td></td>
</tr>
<tr>
<td>Law &amp; Legal Studies (397)</td>
<td></td>
</tr>
<tr>
<td>Library &amp; Information Sciences (58)</td>
<td></td>
</tr>
<tr>
<td>Media &amp; Communication Studies (614)</td>
<td></td>
</tr>
<tr>
<td>Planning &amp; Architecture (330)</td>
<td></td>
</tr>
<tr>
<td>Social Work (170)</td>
<td></td>
</tr>
<tr>
<td><strong>Sciences (4,916)</strong></td>
<td></td>
</tr>
<tr>
<td>Agricultural &amp; Food Sciences (705)</td>
<td></td>
</tr>
<tr>
<td>Biology &amp; Botany (1,038)</td>
<td></td>
</tr>
<tr>
<td>Chemistry (668)</td>
<td></td>
</tr>
<tr>
<td>Computer &amp; Information Science (484)</td>
<td></td>
</tr>
<tr>
<td>Earth &amp; Ocean Sciences (395)</td>
<td></td>
</tr>
<tr>
<td>Mathematics (514)</td>
<td></td>
</tr>
<tr>
<td>Physics &amp; Astronomy (577)</td>
<td></td>
</tr>
<tr>
<td>Resource Management &amp; Forestry (596)</td>
<td></td>
</tr>
<tr>
<td><strong>Social Sciences (6,059)</strong></td>
<td></td>
</tr>
<tr>
<td>Anthropology, Archaeology &amp; Sociology (1,038)</td>
<td></td>
</tr>
<tr>
<td>Economics (369)</td>
<td></td>
</tr>
<tr>
<td>Geography (315)</td>
<td></td>
</tr>
<tr>
<td>Other Social Sciences &amp; Humanities (805)</td>
<td></td>
</tr>
<tr>
<td>Political Science (491)</td>
<td></td>
</tr>
<tr>
<td>Psychology (3,071)</td>
<td></td>
</tr>
<tr>
<td><strong>All Programs (27,397)</strong></td>
<td></td>
</tr>
</tbody>
</table>

Finally, the name of the university of affiliation of doctoral students was written in the form used by Thomson Reuters in the bibliometric database. Table 8 presents a sample record of the final list of doctoral students used for matching scientific publications as well as compiling bibliometric data.
Table 8. Sample record of the GDEU database, including modifications

<table>
<thead>
<tr>
<th>Variable</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students' unique identifier</td>
<td>16162</td>
</tr>
<tr>
<td>Surname</td>
<td>Lariviére</td>
</tr>
<tr>
<td>Given name</td>
<td>Vincent</td>
</tr>
<tr>
<td>Converted name</td>
<td>LARIVIERE-V</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
</tr>
<tr>
<td>University enrolled</td>
<td>McGill</td>
</tr>
<tr>
<td>University name converted</td>
<td>MCGILL-UNIV</td>
</tr>
<tr>
<td>Program enrolled</td>
<td>Information St -T/Ad Hoc</td>
</tr>
<tr>
<td>Domain</td>
<td>Social Sciences and Humanities</td>
</tr>
<tr>
<td>Discipline</td>
<td>Non-Health Professional</td>
</tr>
<tr>
<td>Specialty</td>
<td>Library &amp; Information Sciences</td>
</tr>
<tr>
<td>First semester enrolled in the program</td>
<td>2006AU</td>
</tr>
<tr>
<td>First year enrolled in the program</td>
<td>2006</td>
</tr>
<tr>
<td>Last semester enrolled in the program</td>
<td>2008HI</td>
</tr>
<tr>
<td>Last year enrolled in the program</td>
<td>2008</td>
</tr>
<tr>
<td>Graduating year</td>
<td>-</td>
</tr>
<tr>
<td>Graduating semester</td>
<td>-</td>
</tr>
<tr>
<td>Graduating university</td>
<td>-</td>
</tr>
</tbody>
</table>

3.3. Previous studies on authors’ disambiguation

Given the very high number of doctoral students (N=27,397) for which data on scientific publication had to be compiled, it was decided to use, to a certain extent, algorithms for removing papers authored by homographs. Until recently (Enserink, 2009), the problem of homographs in the scientific literature had not been discussed extensively. The main reason for this is that, in most bibliometric studies, the level of analysis is organizations or geographic entities (universities, provinces, countries, etc.)
and, hence, do not use papers’ author lists. The advent of h-indexes (Hirsch, 2005), aimed at evaluating individual researchers, has increased the interest for individuals’ bibliometric data and, hence, for author disambiguation. That being said, there is still a dearth of information on the extent of the homographic problem in the scientific community. Apart from Aksnes (2008) and Lewison (1996), who respectively compiled data on the extent of homographs among Norwegian researchers and on the frequency of author’s initial letters—but did not test directly their effect on the compilation of bibliometric data on individual researchers—there is very little information on the extent to which researchers share the same name.

Jensen et al. (2008) attempted to compile publication and citation files for 6,900 CNRS researchers using the Web of Science. Instead of removing, in each researchers’ publication file, the papers written by homographs, they evaluated the probability that a given researchers has homographs, and, if this probability was high, they completely removed the researcher from the sample. More precisely, they first measured, by comparing the surname and initials of each researcher (LARIVIERE-V) with some of its variants (LARIVIERE-VG, LARIVIERE-VP, etc.), the probability that the researcher has homonyms. If the researcher had too many variants, it was removed. Their second criterion was related to the number of papers published: if a researcher had too many papers, it was considered as an indication that more than one scientist was behind the records. Hence, researchers had to publish between 0.4 and 6 papers per year to be considered in the sample. Their third criterion was that the first paper of each researcher had to have been published when
the researcher was between 21 and 30 years old. Their resulting database contained 3,659 researchers (53% of the original sample).

This method has at least two major shortcomings. First, the fact that a name (e.g. LARIVIERE-V) is unique does not imply that it represents only one distinct researcher. In this particular case, it could be the surname and initial combination for Vincent Larivièrè, Véronique Larivièrè, Victor Larivièrè, etc. Second, it removes from the sample highly active researchers (who published more than 6 papers per year), which obviously distorts their results. This method is similar to that of Boyack and Klavans (2008), who used researchers with uncommon surnames to reconstitute individual researchers’ publication and patenting activities. Using the combination of the name of the author/inventor and the research institution sign on the paper, they calculated the odds that the paper belonged to the given author. However, contrary to Jensen et al. (2008), they performed some manual validation of their records to measure the error rates.

Another method is that of Han et al. (2005) who, using K-means clustering algorithms and Naïve Bayes probability models, managed to categorize 70% of the papers authored by Anderson-J and Smith-J into distinct clusters. The variables they used were the names of co-authors, the name of journals and the title of the papers. The assumption behind this algorithm is that researchers generally publish papers on the same topics, in the same journals and with the same co-authors. A similar method was also used by Torvik et al. (2005) using Medline. Similarly, Wooding et al.

---

28 The bibliometric part of their paper used the Scopus database, which, contrary to Thomson Reuters’ databases, links names of authors with institutional addresses.
(2006) used co-authors for removing homographs from a sample of 29 principal investigators funded by the Arthritis Research Campaign. For each author, they first found a core of papers which, without a doubt, belonged to the right researcher. Using this ‘core’ subset of papers in the specialty of arthritis, they created, for each researcher, a list of co-authors which were used to gather papers in areas other than arthritis. A novel aspect of this study is that several rounds of co-author inclusion were performed, increasing between each round the number of co-authors in the core. After three rounds of the algorithm, 99% of the authors’ papers were assigned—which could be considered as the recall of papers—with only 3% of false positives (97% precision). This method is very similar to that used by Kang et al. (2009).

Finally, Aswani et al. (2006) used, in addition standard bibliographic information (abstract, initials of the authors, titles and co-authors), automatic web-mining for grouping papers written by the same author. The Web-mining algorithm searches for the full names of the authors, finds their own publications’ page, etc. Their results show that web-mining improves the clustering of papers into distinct authors, but the small sample used in the study makes the results less convincing.

The methods presented in these studies indicate that it is possible, using regularities in the publication patterns of researchers, to cluster papers authored by the same author. However, none of these methods can be used as such, since the full population of doctoral students—and not a sample—is analyzed in this thesis and a low error rate is needed, especially in terms of false positives. Along the same lines,
contrary to these studies, this thesis benefits from having a list of distinct researchers (students or university researchers) with their full names and affiliations, which greatly helps the removal of papers authored by homographs. This thesis thus propose a new method for removing papers authored by homographs, which draws on the patterns found in the dataset of the manually-cleansed publication files of the full population of professors and university-affiliated researchers in the province of Quebec for the period 2000-2007. The regularities found in this dataset are presented in the following section.

3.4. Publication patterns of Quebec’s university-affiliated researchers

Studies presented in the preceding section have an *a priori* design, in the sense that they assumed that several patterns could be found in researchers’ publications, and then empirically tested these patterns to cluster papers. Instead of assuming the existence of such patterns, the method used in this thesis is to first establish the existence of such patterns, using the cleaned publication files of the full population of Quebec university professors and researchers for the period 2000-2007 (N=13,479) (Gingras et al. 2008; Larivière et al. 2009)—the same period used in this thesis.

Indeed, since 2005, the OST has been doing, within the framework of research contracts with Quebec’s MDEIE and three research councils (FRSQ, FQRNT and FQRSC), the reconstitution of the publication files of each university professor and researcher. This dataset serves as the backbone for finding the relationships between the disciplinary affiliation of university researchers and the discipline of their
publications. The patterns found in this dataset are presented in this section and serve as an input for the algorithmic attribution of publications to doctoral students. The difference between university researchers’ papers manually assigned and rejected allows the testing of the algorithm. This method fairly assumes that patterns found in professors and university-affiliated researchers are similar to those of doctoral students.

The database on university researchers’ papers was obtained by matching the surname and initials of these researchers to the surname and initials of authors of Quebec’s scientific articles indexed in OST’s version of the Web of Science. This first match resulted in a database of 125,656 distinct articles and 347,421 author-article combinations. Each article attributed to each researcher was then manually validated in order to remove the papers authored by homographs. This essential but time-consuming step reduced the number of distinct papers by 51% to 62,026 distinct articles and by 70% to 103,376 author-article combinations. Analysis of this unique dataset, including the characteristics of both assigned and rejected papers, sheds light on the extent of homonyms in Quebec’s scientific community.

In order to test the reliability and reproducibility of the manual validation of university researchers and professors’ publication files, tests with different individual ‘attributors’ were performed for a sample of 1,380 researchers (roughly 10% of the researchers). It showed that for most publication files, the two coders manually assigned exactly the same papers. More specifically, 1,269 files (92%) researchers had exactly the same papers assigned. A difference of one paper was found in 72 cases
(5.2%), 2 papers in 15 cases (1.1%), 3 papers in 9 cases (0.7%) and 4 papers in 3 (0.2%) cases. The remaining 12 files had a maximum difference of 12 papers each. In terms of author-article combinations, the error rates are even lower. Out of the 12,248 author-article links obtained the first time, 12,124 (or 99%) remained unchanged the second time. Manual validation is thus quite reliable and reproducible.

The first interesting piece of information found in this dataset is the percentage of papers of each researcher that were retained after manual validation. More specifically, these cleaned publication files made it possible to estimate the extent of homographs problems for all Quebec university researchers for whom at least one article was automatically matched (N= 11,223) using the name of the researcher within papers having at least one Quebec institutional address29. With an automatic matching of researchers’ names, compared to a cleaned publication file (Figure 5):

- The papers matched for 2,972 researchers (26.5%) were all rejected which, in turn, meant that they had not actually published any papers (all papers were written by homographs);
- Between 0.1 and 25% of the papers matched were assigned for 1,862 researchers (16.6%);
- Between 25.1 and 50% of the papers matched were assigned for 975 researchers (8.7%);
- Between 50.1 and 75% of the papers matched were assigned for 722 researchers (6.4%);
- Between 75.1 and 99.9% of the papers matched were assigned for 818 researchers (7.3%);

---

29 Thus, 2,256 of Quebec’s researchers did not publish any paper during that period nor had any of their homographs.
- The papers matched of 3,874 researchers (34.5%) were all conserved after manual validation (i.e., they had no homographs within the subset of Quebec papers).

**Figure 5. Percentage of papers assigned after manual validation, by researcher**

A ‘rough’ matching without removing papers authored by homographs is thus valid for slightly more than a third of the researchers. On the other hand, the scientific production of the remaining two-thirds was significantly overestimated. Since it is impossible to know *a priori* which researchers will be overestimated and which will not, the validation each paper from each researcher is, theoretically, needed. Papers of these publication files were all manually validated (assigned or rejected) and serve, in a reverse engineering manner, as a test bed for finding patterns in the publications of researchers. These patterns will, in turn, allow the creation of an algorithm for the automatic attribution of publications to doctoral students, as there is no reason to
believe that doctoral students’ publication patterns are different from those of researchers.

The first step for finding regularities in university researchers’ publication patterns was to assign a CIP disciplinary code to each of the researchers’ department. In order to be able to use these patterns for the analysis of doctoral students’ publications, the same disciplinary classification was used (see section 4.3.1). In a manner similar to that of Wooding et al. (2006) for arthritis research, papers were then analyzed in order to find characteristics which could help isolating a core of papers for each researcher—i.e. a subset of all of each researcher’s papers that we are sure are not those of homographs. This was more complex in the context of this thesis, as the core papers had to be found for researchers that could be active in any field of science and not only in arthritis. After several rounds of empirical analysis, the combination of three variables optimized the ratio between the number of papers found and the percentage of false positives. Figures 6 and 7 present the two sets of criteria with which a core set of papers could be found for university-based researchers. Figure 6 presents the first matching criteria: the complete name of researchers matched with the complete name of authors (available in the Web of Science since 2006) and the name of the researcher’s university matched with the name of the university on the paper.
Figure 6. First matching criteria for creation the core of papers

- Full name of the researcher =/≈ Full name of the author
- University of the researcher = Institution/Affiliation of the author

Example:
- Lanvière, Vincent
- MCGILL-UNIV

Figure 7 presents the second matching criteria. Firstly, the name of the author of the paper had to be written exactly in the same manner as the name of the researcher in the list. Secondly, the institution appearing on the paper (or its affiliation, e.g. Royal Victoria Hospital is affiliated to McGill University) had to be the same as the institution appearing on the list and, thirdly, the discipline of the journal in which the paper is published, the department or the institution of the authors had to be similar\(^{30}\) to the department of the researcher as it appeared on the list of university professors and researchers or the discipline of the paper had to be among the five disciplines in which researchers from this department published.

\(^{30}\) The similarity threshold was set at 25% using Microsoft SQL Server.
Following Boyack and Klavans (2008), an analysis of rare surnames was also performed, which were defined as surnames only belonging to one individual in the list of university researchers. Hence, all papers authored by researchers having a rare name, and on which their institution of affiliation appears, were included in core papers. In Table 9, these three criteria allow the creation of a core of papers for more than 75% of the individual researchers for which at least one paper has been manually assigned (8,081), matches 56.4% of their distinct papers and 47.5% of the author-paper combinations, e.g. LARIVIERE-V and paper ‘X’. At each level of analysis, the number of false positives is rather low; and is especially low at the level of author-paper combinations (less than 1%).
Table 9. Results of the matching of core papers at the levels of university researchers, articles and author-paper combinations

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Manual validation (N)</th>
<th>Automatic creation of the core N</th>
<th>%</th>
<th>False positives N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>8 081</td>
<td>6 117</td>
<td>75,7%</td>
<td>344</td>
<td>4,3%</td>
</tr>
<tr>
<td>Articles</td>
<td>62 629</td>
<td>35 353</td>
<td>56,4%</td>
<td>772</td>
<td>1,2%</td>
</tr>
<tr>
<td>Author-paper combinations</td>
<td>97 850</td>
<td>46 472</td>
<td>47,5%</td>
<td>809</td>
<td>0,8%</td>
</tr>
</tbody>
</table>

Another set of regularities was found in individual researchers’ publication patterns. The idea behind this search for patterns for individual researchers was to be able, using a subset of core papers, to find other papers that belonged to the researchers but that did not exhibit the characteristics found in Figure 6 and 7. To do so, each researcher’s publication record was divided into two distinct 4-year time periods: 2000 to 2003 and 2004 to 2007.

Using the characteristics of the papers published by each given researcher during the first time period, an attempt was then made to automatically assign to the same researcher the papers published during the second time period. Two indicators were quite successful in doing so: 1) the use of the same words in the title, author keywords and abstract fields of upcoming publications (Figure 8) and 2) the citing of the same references (Figure 9) of papers for which the Thomson name [e.g. LARIVIERE-V] and the institution [MCGILL-UNIV] also matched. Figure 4 presents the percentage of rightly and wrongly attributed papers, as a function of the keyword index. The keyword index is a simple indicator compiled for each 2004-
2007 paper matched to a researcher, based on the keywords of the papers assigned to the researcher for the period 2000-2003. Its calculation is as follows:

\[
K_{\text{pr}} = \left( \frac{N_{\text{kpm}}}{N_{\text{kp}}} \times \frac{1}{\sqrt{N_{\text{kt}}}} \right) \times 100
\]

Eq. (1)

Where \( N_{\text{kpm}} \) is the number of keywords of a 2004-2007 paper that match the keywords used in the 2000-2003 papers of a researcher, \( N_{\text{kp}} \) is the total number of keywords of the 2004-2007 paper and \( N_{\text{kt}} \) is the total number of keywords used in all the 2000-2003 papers assigned to the researcher. The square root of \( N_{\text{kt}} \) was used instead of \( N_{\text{kt}} \) alone in order to obtain an overall number of keywords that is not too high—especially for very productive researchers. The result is multiplied by 100 in order to be closer to an integer.

Figure 8 shows that when the keyword index is at two, about 90% of the papers rightly belong to the researcher and that slightly greater than 10% are false positives. When the keyword index is greater than 2 (3 or more), the percentage of rightly assigned papers rises above 95%, and stays at this level until the keyword index reaches 7, where about 100% of the papers are assigned to the right researcher.

These numbers mean that it is possible to rightly assign papers to a researcher using the regularities found in the title words, keywords and words of the abstract.
Figure 8. Percentage of rightly assigned and wrongly assigned papers, as a function of the keywords previously used by a university researcher, 2000-2003 and 2004-2007

Figure 9 presents the references index for 2004-2007 papers, based on papers published between 2000 and 2007. The references index is very similar to the keyword index previously presented; it is based on the pool of references made previously (2000-2003) by the researcher. Its calculation is as follows:

\[ R_{i_{pr}} = \left( \frac{N_{rpm}}{N_{rp}} \times \frac{1}{\sqrt{N_{rt}}} \right) \times 100 \]  

Eq. (2)

Where \( N_{rpm} \) is the number of references of a 2004-2007 paper that match the references used in the 2000-2003 papers of a researcher, \( N_{rp} \) is the total number of references of the 2004-2007 paper and \( N_{rt} \) is the total number of references used in all the 2000-2003 papers assigned to the researcher. Again, the square root of \( N_{rt} \) was used instead of \( N_{rt} \) alone in order to have an overall number of cited references that is not too high. The result is also multiplied by 100 in order to be closer to an integer.
It shows that as soon as a signal is obtained, i.e. that at least one of the referenced work of the 2004-2007 paper was previously made in the 2000-2003 dataset, more than 90% of the papers rightly belong to the researcher. When the references index increases to 1 or above, the quasi-totality of the papers rightly belong to the researcher.

**Figure 9. Percentage of rightly assigned and wrongly assigned papers, as a function of the references previously made by a university researcher, 2000-2003 and 2004-2007**

When the articles matched using these two indicators (set at 2 or more for the keyword index and at >0 for the references index), 10,892 additional papers were assigned, with only 236 papers being false positives (2.2% of the added papers), for an overall error rate at the level of papers of 2.2% and of 1.7% at the level of author-paper combinations (Table 10). Since this second round can only be made for researchers for which a certain number of core papers were matched, the number of
researchers stays the same, but slightly more researchers have at least one paper wrongly assigned (6.7%).

Table 10. Results of the matching of core papers and papers with the same keywords or cited references, at the levels of university researchers, articles and author-paper combinations

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Manual validation (N)</th>
<th>Automatic creation of the core</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Researchers</td>
<td>8 081</td>
<td>6 117</td>
<td>75,7%</td>
</tr>
<tr>
<td>Articles</td>
<td>62 629</td>
<td>46 245</td>
<td>73,8%</td>
</tr>
<tr>
<td>Author-paper combinations</td>
<td>97 850</td>
<td>64 765</td>
<td>66,2%</td>
</tr>
</tbody>
</table>

I then proceed to a second round of automatic matching of papers with the same references and keywords, using only the Thomson name [e.g. LARIVIERE-V] and the province [QC], but not the institution [MCGILL-UNIV]. Using this method, 3,645 additional papers were retrieved, of which 674 were false positives (Table 11). Although this percentage seems quite high, the overall proportion of false positives at the level of articles remains quite low (3.2%) and is even lower for author-paper combinations (2.3%).
Table 11. Results of the matching of core papers and papers with the same keywords or cited references, without the ‘same institution’ criteria, at the levels of university researchers, articles and author-paper combinations

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Manual validation (N)</th>
<th>Automatic creation of the core</th>
<th>False positives</th>
<th>N</th>
<th>%</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>8 081</td>
<td>6 117</td>
<td>576</td>
<td>94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Articles</td>
<td>62 629</td>
<td>49 890</td>
<td>1 577</td>
<td>32%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author-paper combinations</td>
<td>97 850</td>
<td>72 918</td>
<td>1 682</td>
<td>23%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to increase the number of researchers for which a certain number of core papers could be found, the relationship between the discipline of researchers and the specialty of papers was analyzed. For each of the 5,615 existing combinations of disciplines of departments (as defined in Table 7) and of disciplines of publications (as defined in Tables 2 and 3), a matrix of the percentage of papers from each discipline of publication that rightly belonged to researchers from each department was calculated. It was found that papers published in the main specialty in which researchers from a given specialty publish were more likely to belong to the right researchers. For example, 100% of the 186 papers published in geography journals that matched the names of authors of geography departments belonged to the right researcher. The same is true for several other obvious department-specialty relationships, such as university researchers from chemical engineering departments publishing in chemical engineering journals (99% of the 1,017 papers rightly assigned), but also for less obvious relationships such as researchers in civil engineering publishing in Earth & planetary science journals (95% of the 316 papers rightly assigned).
On the other hand, all the 333 papers published in biochemistry & molecular biology journals that matched authors from the disciplines of anthropology, archaeology & sociology belonged to the wrong researcher. The same is also true for the 202 papers published in organic chemistry that matched authors from business departments. Given that no university-affiliated researcher from this domain has ever published in journals of this specialty, there are low chances that a doctoral student of the same domain has done so.

Figure 10 presents the matrix of the percentage of assigned papers, for each combination of discipline of departments (x-axis) and specialty of publication (y-axis). Darker zones are combinations of specialties of publications and disciplines of departments where a larger proportion of papers were accepted during manual validation; lighter zones are combinations where a majority of papers were rejected during manual validation. This figure illustrates that there is a majority of discipline of department/specialty of publication combinations where the quasi totality of papers were authored by homographs (light zones), and a few darker zones where a large proportion of papers were accepted. In other words, the purpose of this figure is not to provide the reader with the specific combinations of departments/specialty of journals where all papers were accepted—for example, researchers from departments of information science & library science publishing in journals of library & information sciences—or rejected, but rather to present the ‘landscape’ of the relationships between the specialties of journals and the disciplines of departments.
The presentation of this landscape clearly shows that there are some combinations where the majority of papers were assigned during manual validation and others where only a minority of papers was assigned during the process. Our algorithm thus focuses on these light zones to automatically exclude doctoral students’ papers from a given department published in given specialty, and on dark zones to automatically include papers from other department/specialty combinations.
Figure 10. Percentage of papers assigned after manual validation, for each combination of discipline of departments (x-axis) and specialty of publication (y-axis).

Figure 11 aggregates, by rounded percentage of properly attributed papers, the numbers of rejected and of accepted author-paper combinations. We see that the number of wrongly assigned papers drops significantly for department/specialty combinations greater than 80%, and even more after 95%. These percentages were
thus used to automatically assign papers in the specific discipline that matched researchers from given departments. In order to reduce the number of false positives, a 95% attribution rate was chosen for papers on which the institution of the author does not appear (but only the province). This includes a total of 17,002 papers, of which 16,518 are properly attributed and only 484 are inaccurately attributed (2.8%). For papers on which the institution of the researcher appears, an 80% attribution rate was used. This attributed 68,785 papers, of which 10.7% were false positives.

Figure 11. Number of rightly and wrongly assigned author-paper combinations, as a function of attribution percentage of papers from a discipline to authors from a department

One must note that all these processes were performed in parallel; a paper assigned with one of these criteria could have been already attributed during another step of...
the matching. Hence, the numbers of papers presented here include several papers that were already matched using one of the criteria previously presented in this section. Table 12 presents the error rates for all of the steps combined. The inclusion of the algorithm based on the matrix of departments and disciplines of publication added 310 researchers in the subset of those with at least one paper in the core. On the whole, the multiple algorithms used so far automatically attributed papers for 6,427 researchers, for a total of 50,353 papers and 73,331 author-paper combinations.

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Manual validation (N)</th>
<th>Automatic creation of the core</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Researchers</td>
<td>8 081</td>
<td>6 427</td>
<td>79,5%</td>
</tr>
<tr>
<td>Articles</td>
<td>62 629</td>
<td>50 353</td>
<td>80,4%</td>
</tr>
<tr>
<td>Author-paper combinations</td>
<td>97 850</td>
<td>73 771</td>
<td>75,4%</td>
</tr>
</tbody>
</table>

All the patterns presented so far in this section allow the creation of a dataset of papers that are likely to belong to the right researcher and assign at least one paper to almost 80% of the researchers. They make possible the creation of a core set of papers, as well as of a few other layers of papers, based on the similarity of their characteristics to those included in the core. The following algorithm does the opposite and aims at finding indications that the paper clearly does not belong to the researcher.
As shown on Figure 11, there are several combinations of discipline of departments and specialties of papers where the vast majority of papers were rejected during manual validation. Indeed, if no university researcher from the department X has ever published in the specialty Y, no doctoral student is likely to do so. Papers falling into these combinations could thus automatically be rejected. Doctoral students are indisputably less productive than established researchers and, hence, there are few chances that their scientific production is spread across more specialties than that of university researchers.

These patterns not only allow the rejection of papers, but also permit the closing researchers’ publication files, as all of their papers can either all have been assigned—using the methods previously presented—or rejected using the department/specialty matrix. Using a 50% threshold, 202,928 author-paper combinations are rejected, of which 183,656 are real negatives (91%), and only 19,272 are false negatives (9%). These rejected author-paper combinations account for a significant share (90%) of all rejected combinations (226,325).

After all these steps, 5,036 publications files out of the 13,479 (37.4%) were automatically marked as closed (including the 2,256 files for which no paper, either authored by a researcher in the list or by a homograph), as all of their papers were either all assigned or all rejected. Another 6,069 researchers had at least one of their papers automatically assigned (45%), for a total of 50,353 papers, with 1,633 being false positives (3.2%). On the whole, this algorithm provides attribution information on at least one paper for 11,105 (82.4%) out of all 13,479 researchers, or on 8,849
out of the 11,223 researchers (78.8%), when one excludes the 2,256 files for which no paper matched, either authored by the researcher or a homograph.

Hence, there are still 2,374 researchers for which no automatic decision on any of their matched papers can be made (attribution or rejection) and, hence, for which a complete manual validation needs to be performed. This algorithm can nonetheless be very helpful, as it automatically assign a large proportion of papers, excludes an even larger one and reduces from 11,223 to 2,374 (79%) the number of researchers for which a complete manual validation has to be performed. That being said, the application of this algorithm might not be entirely straightforward. Students’ smaller scientific production—one paper—will probably make it more difficult to use previously published papers as indicators of the likeliness that another paper is theirs. Along the same lines, we can expect doctoral students to be more mobile than established researchers and, hence, sign papers without the name of the university where they are doing their PhD (e.g. papers written during their Masters degree in another university but published once they are enrolled in a doctoral program).

3.5. Attribution of publications to doctoral students

This section details the process by which publications were assigned to doctoral students. It first presents the results of the application of the automatic attribution algorithm based on university researchers’ and professors’ publications patterns presented in the preceding section, to doctoral students’ scientific publications. As one might expect, not all doctoral students’ publication files could be processed using the algorithm. Secondly, it presents the method and interface used for the
manual validation of the remaining publication files. Finally, the last section discusses the overall limitations of the methods used.

### 3.5.1. Algorithmic attribution

Matching the names of the 27,397 doctoral students with the names of authors of Quebec papers using their surname and initials yielded 313,367 author-article combinations for 15,702 doctoral students. Hence, 11,695 doctoral students had not contributed to any paper, and nor had any of their homographs over the 2000-2007 period. As one can expect from the patterns found in university researchers’ publication files, a vast majority of these author-article combinations are false positives.

Out of these 15,702 doctoral students who matched with at least one paper, the application of the first step of the algorithm (Figure 6 and 7) managed to assign at least one paper for 5,621 students, for a total of 12,715 papers. In a second step, using the references and keywords used in the papers, as well as the department/specialty matrix at 80% and 95% (Figure 11), 11,556 additional papers were automatically assigned to these researchers, for a total of 24,271 papers.

Application of the matrix also added 1,508 students in the process, for a total of 7,129 students. Out of these, 4,243 students’ files were automatically closed as all of their papers—generally one or a few—were assigned.

On the other hand, the application of the same department/specialty matrix at 50% for automatically removing author-article combinations allowed the rejection of
34,079 articles and, thus, automatically closed 5,352 publication files, as all of their papers were rejected. Globally, the application of the automatic attribution algorithm allowed the closing of 9,595 files out of the 15,702 for which at least one author matched the name of a doctoral student. To this were added the 2,886 files for which at least one paper was automatically assigned. Hence, on the whole, paper attribution information was obtained for 12,481 students out of 15,702 (79.5%), which is a percentage very similar to that obtained for university researchers. This percentage rises to 88% (24,172 out of 27,393) if one includes the 11,691 students for which no paper matched, either authored by the doctoral student or a homograph.

### 3.5.2 Manual validation

After completion of this automatic attribution process, a significant number of doctoral students’ publication files still had to be validated manually (6,107 doctoral students, 22.3% of the population). Of this group, 2,886 PhD students already had at least one paper automatically assigned by the algorithm.

Figure 12 presents the interface built for the manual attribution of papers, used for Quebec university researchers’ and professors’ publication files as well as for those of doctoral students. For each doctoral student (or university researcher), a list of papers having at least one institutional address from Quebec matching the name of the student is shown. For each of these papers preliminarily matched, it provides a unique identifier, its title, publication year, surname and initials of the author who matched the name of the researcher, complete name (when available), institution and department appearing on the paper, and the discipline and specialty of the journal
where the paper was published. The same paper can be duplicated several times, as more than one Quebec institution might appear on the paper. At the bottom of the interface, the following information on the student appears: complete name, institution and program name.
Figure 12. Interface for the manual attribution of publications to doctoral students
Manual validation is thus performed by rightly assigning the paper-institution combination that belongs to the student. This manual validation is generally made by searching the title of each of the papers on Google to find their electronic versions on which, generally, the complete names of the authors are written. This often helps to decide if the papers belong to the researcher. Another method is to search the name of the student on Google to find his/her website to get an indication of is publications’ list or CV. After a few papers, one generally understands the publication pattern of the student and correctly attributes his/her papers.

These assigned papers are highlighted in green and other duplicates of the assigned paper with other institutional addresses from Quebec automatically become red, as the same papers cannot be assigned twice to the same researcher. For obvious reasons of internal logic, the same author-institution-paper cannot be assigned to more than one researcher. However, as papers are often written in collaboration, the same institution-paper combination can be assigned to another student with a different name. Papers that do not belong to the researcher are left unassigned—no command has to be performed. Once all the papers that belong to the student are assigned, a manual command is performed to close the student’s publication file.

Once the manual validation of the remaining cases was completed, another round of manual validation was performed, in which a different individual was asked to go over 100 random cases in order to mistakes in either the manual or the automatic attribution. The overall attribution process proved to be quite reliable, as differences in paper attribution were found in only one case through this double verification.
process. Finally, a last round of validation was performed for doctoral students’ publications files having more than 50 publications. This last step caught 6 doctoral students’ files wrongly assigned—all in high energy physics, where author lists often contain hundreds of names—for 456 author-article combinations.

After all of these steps, 31,738 author-article combinations were retained (out of the 313,367 originally obtained) for 25,159 distinct papers and 8,468 doctoral students. Manual validation thus added 1,339 students31 and 888 papers to the dataset. Numbers presented in chapters 4 and 5 are a subset of this dataset, as they only take into account the papers published during doctoral studies.

3.5.3. Limitations to the methods

In addition to the well-known limitations of bibliometric methods and databases presented in chapter 2, there are also limitations to the results obtained with our method for assigning articles to individual researchers. Most of these limitations may increase the rate of false negatives, i.e. papers that were indeed written by PhD students but that were nonetheless excluded from our dataset.

First, doctoral students’ were retrieved from the subset of WOS papers on which there is a research organization located in Quebec. Hence, students who obtained their bachelor or masters’ degree in another province or country, or who are from

31 The number of students is higher than the number of papers because a certain number of papers automatically assigned were rejected during the manual validation. These papers were mainly in the field of high energy physics. This specialty had to be removed from the algorithm because its papers have a very high number of authors which, in turn, increases the probability of homographs.
other Canadian provinces or other countries might have their scientific production underestimated. Indeed, they might have published papers prior to their admission in a PhD program in Quebec and these papers are less likely to have a Quebec address, unless they were written in collaboration with Quebec researchers. This limitation might also affect Quebec students who have been performing part of their PhDs in a foreign country and have signed papers within the foreign research institution to which they were affiliated during their stay. Along the same lines, if for some reason the name of the Quebec research institution where the research was conducted does not appear on the paper—be it a mistake from the author, the journal or from Thomson Reuters’ indexing—the publication will not be included in the study.

A second set of limitations might come from a change in doctoral students’ surnames. For instance, if a doctoral student has changed his/her surname from Smith to Johnson between enrolling in a program and the publication of papers, the papers will not be retrieved. That being said, changes in surnames—especially because of marriage—are less common in Quebec, so this might not affect our analysis. Along the same lines, a change in students’ given names might also affect the resulting data. If for some reasons a young researcher changes his/her given name during his/her enrolment in the PhD or authors scientific papers with a name different from his/her legal name (‘Bob’ instead of Robert, etc.), then such papers will not be included in the study because the attribution of papers is based on the initial of the given name.
Another potential limitation—which is difficult to assess, however—is that doctoral students could be more ‘interdisciplinary’ than more established researchers, as they might be more prone to change their research topic during the course of their masters or doctorate degrees. Hence, the keywords and or cited references indexes might be effective for finding subsequent papers belonging to the doctoral student. Along the same line, a higher level of interdisciplinarity in doctoral students could also affect the assumption that doctoral students’ patterns are similar to those of established researchers. That being said, given that most doctoral students’ papers are written in collaboration with faculty (Figure 19), it is not very likely that their papers exhibit a pattern different from that of established researchers.

In sum, the data provided in this section clearly show that our algorithm and the subsequent manual validation effectively accomplish the retrieval of Quebec doctoral students’ papers. Since these limitations are more likely to create a certain number of false negatives than false positive, the data analyzed in this thesis can be considered as the minimum number of papers to which students have contributed.

### 3.6. Bibliometric indicators

This thesis uses several bibliometric indicators to analyze and compare doctoral students’ publications. These indicators are compiled for 1) doctoral students 2) university-affiliated researchers and professors and 3) other Quebec papers to which doctoral students have not contributed. The database on university-affiliated researchers and professors used in this thesis—and on which is based the automatic attribution of doctoral students’ publications presented in section 3.4—is mainly
composed of university professors, but also includes other university-affiliated researchers (clinicians, etc.). Hence, for the sake of conciseness, these data are often referred to as ‘faculty’ or ‘professors’ publications.

3.6.1 Number of publications

In this thesis, publication counts are based on the number of articles, notes and review articles to which authors from Quebec contributed. Hence, editorials, book reviews, letters to the editor or meeting abstracts are excluded from the analysis, as they are not generally considered as original contributions to scholarly knowledge (Moed, 1996). These numbers are based on full counting of papers, as opposed to fractional counting sometimes used in bibliometrics. Hence, each individual or organization contributing to a paper is assigned one ‘full’ contribution, instead of a fraction of a contribution, irrespective of their rank in the author order.

For the purpose of brevity, publications to which doctoral students have contributed are sometimes referred to as ‘doctoral students’ publications’, even though they are rarely the sole authors of the papers on which they are authors. The same can also be said of papers to which faculty members contributed or of Quebec papers which are, in this thesis, sometimes referred to ‘faculty papers’ or ‘Quebec papers’, respectively, and of which neither faculty members, nor ‘Quebec researchers’ are always the sole authors.

Doctoral students do not keep this status eternally: they either graduate or leave the program without graduating. In this thesis, papers are considered as being authored
by doctoral students when at least one of the authors is enrolled in a PhD program in one of Quebec’s universities during the publication year of the paper or has been enrolled during the year prior to the publication of the paper. In other words, in line with Lee (2000), doctoral students’ papers are still considered as such until one year after they graduate or leave the program. Although Anwar (2004) exposes that a majority of papers drawn from the doctoral thesis were published during the two years following graduation year, it was decided to keep only a one year lag, as a two year lag would necessarily include several papers written after graduation. Note that this practice is also found in the large teams of particle physics, where researchers get to sign the papers coming out of the experiment until one year after they leave the team (Biagioli, 2003). Moreover, the goal of this thesis is not to analyze scientific papers drawn from theses, but rather to analyze the scientific production of doctoral students during their studies.

3.6.2 Average of relative impact factor

Created in the 1960s, the impact factor is an indicator that, in recent years, has received much attention in the scientific community (Archambault and Larivière, 2009). Indeed, this indicator of the impact of scientific journals—originally used by librarians for choosing which scientific periodicals to subscribe to—gradually became used as an indicator for evaluating research and researchers. Thomson Scientific compiles each year the impact factor of journal it indexes in the SCI and SSCI (Thomson Reuters, 2009). Its calculation is as follows:

\[ IF_{jy} = \frac{C_{jy}}{P_{jy-1} + P_{jy-2}} \]  

Eq. (3)
Where \( IF_{jy} \) is the impact factor of a journal \( j \) for a given year \( y \), \( C_{jdy} \) is the number of citations received by the journal in year \( y \) for all documents \( d \) published in the journal during the two years prior to year \( y \) and \( P_{jdy-1} + P_{jdy-2} \) is the number articles, research notes and review articles published in the journal \( j \) during the two years prior to year \( y \) by the same set of documents \( d \). The impact factor is, thus, an indicator of the average short-term impact of papers published in a given journal. Despite its wide use in research evaluation, the impact factor thus calculated is not without problems. Five limits are generally attributed to the impact factor (Archambault and Lariviere, 2009).

First, there is an asymmetry between what is included in the numerator and what is included in the denominator. Indeed, while citations received by all types of documents published by the journal (articles, notes and reviews, but also editorials, letters to the editor, news items, etc.) are counted in the numerator, only articles, notes and review articles—considered as citable documents—in the denominator. Thus, a journal that publishes many editorials and letters to the editor that are cited increases its impact factor compared to journals that only publishes articles. This is particularly the case of journals *Nature* and *Science*. Although the extent of this practice is not quantified some publishers have also been known to artificially increase the impact factor of their journals by changing the names of certain type of documents from articles (citable documents) to letters (non citable documents), thus removing them from the denominator. This is corrected in our own compilation of the impact factor (Eq. 3).
Secondly, the probability of an article being cited is not the same in all disciplines. For example, the average impact factors of journals in biomedical research are much higher than in chemistry or physics. This is obvious given that the average number of references per article is much higher in biomedical research than in these two other disciplines. One can also observe a marked difference between different specialties within a given discipline.

Thirdly, the impact factors of journals include journal self-citations, that is, citations made to articles published in the same journal. Although the inclusion of such self-citations poses no problems \textit{a priori}, it began to become more problematic when some journal publishers started to suggest to authors to cite papers previously published in the same journal.

Fourthly, the ‘observation’ window during which the citations are counted (two years) is generally considered too short to measure the overall impact of the articles. In certain disciplines where the pace of research is slower—such as those of the social sciences—scientific documents need several years before receiving a significant number of citations. Finally, given that the distribution of citations received by articles published in a journal is non-parametric (Seglen, 1992)—a minority of articles receiving the vast majority of citations—the impact factor of the journal is a bad predictor of the number of citations received by the articles.
Impact factors compiled in this thesis correct most of these limitations. First, the impact factors of each journal indexed in the database were recompiled by 1) correcting the problem of the numerator and denominator—only citations received by articles, notes and review papers are counted—and 2) using citation windows of 2 years, 5 years and 10 years. Given that the citation window of 2 years is sufficient for the vast majority of research in the medical field—as the pace of publications and citations is fast enough—this citation window is used for these disciplines. For the social sciences, a citation window of 5 years was used and for the humanities, a citation window of 10 years, given that the pace of publication and citation is much slower in these disciplines.

As a second step, in order to make data comparable between different specialties, each value of impact factor has been normalized by the average impact factor of journals in the same specialty (Schubert and Braun, 1986; Moed, De Bruin, and van Leeuwen, 1995). Hence, if an article is published in a journal with an impact factor of 2 and the average impact factor of his specialty is 4, it will obtain a value of relative impact factor of 0.5. These relative impact factors values are then averaged for the group of researchers studied to generate the average of relative impact factor (ARIF) indicator. When the ARIF is greater than 1, the articles of the group of researchers are, on average, published in journals with impact factors higher than the world average in the fields studied. Conversely, an ARIF below 1 means that the impact factors of journals where the researchers have published are below the world average.
Journal self-citations were kept in the calculation of the ‘corrected’ impact factor because 1) otherwise than through anecdotal evidence, we do not know the extent of the editors’ self-citation policies and 2) this could disadvantage journals of smaller specialties in which the majority of the scientific production of the area is concentrated, as these journals will indisputably have high self-citation rates. This is, for instance, the case of the journal *Scientometrics*.

Finally, the fact that the impact factor has a relatively low predictive value on the number of citations received by articles is not necessarily a problem, since we do not consider the ARIF as a predictor of the number citations to be received by articles, but rather as a measure of the impact of scientific journals and of their prestige. For instance, it has been recently shown that a journal’s impact factor has an impact on the number of citations received by articles (Larivière, and Gingras, 2010). In this thesis, however, the observed impact of articles is being measured by the citations actually received by the articles.

Finally, there is a need mention that the impact factor has nevertheless an advantage over the citation received. Indeed, since it is based on articles previously published in the journal, it is available upon publication of the article. To the contrary, one has to wait several years for citations to accumulate and, therefore, to be measured.

### 3.6.3 Average of relative citations

Citations received by articles are generally regarded as a fair measure of their impact. These measures of citations have nonetheless several limitations. First, such as the
impact factor, citations cannot be compared across disciplines and specialties because of different referencing patterns. In this thesis, this limit is corrected by the normalization of each article’s citation count by the number of citations received by articles published the same year in journals of the same specialty. These normalized citation scores are then aggregated to form the average of relative citations (ARC).

Secondly, we must allow some time for citations to accumulate. In other words, at their moment of publication, articles obviously have not received any citations—yet. These citations will accumulate—at a fast or slow pace, depending on the discipline—in the months and years following publication. In order to take into account the different pace of research of the different fields, an open citation window was used. More precisely, this means that the total number of citations received since publication by a given paper are counted, and then divided by the average number of citations received by papers published the same year. For example, citations received by a paper published in 2001 are counted for 6 years (in addition to publication year), and then divided by the average number of citations received during the same period by papers of the same specialty also published in 2001. This citation window is probably not long enough to measure research impact in the humanities, as slightly more than 15% of the papers in these disciplines have received at least one citation 5 years after publication (Larivière, Gingras and Archambault, 2009).

Thirdly, citations are probably the indicator on which researchers have the most influence because they can cite themselves (self-citations). This is less likely to
influence doctoral students’ citation rates, as they have less previous publications to cite in their actual publications. That being said, citation measures presented in this thesis exclude first author self-citations. Finally, note that the citations are only counted for publications indexed in the database; citations received by books or conference proceedings published are not included.

3.6.4 Percentage of papers in the top 5% of cited papers
This indicator is a measure of the percentage of papers of a group of researchers that are among the most cited papers of their specialty, as calculated by the ARC. If a group of researchers (students, university researchers, etc.) has more than 5% of their papers in the top 5% of most cited papers, then it means that they produce more highly-cited papers than expected. If this percentage is below 5%, it means that they produce less high impact papers than expected.

3.6.5 International collaboration
A Canadian article is considered to be the result of an international collaboration when at least one of the institutional addresses appearing on the paper is from a foreign country. This indicator is generally presented as a percentage (percentage of papers written in international collaboration).

3.6.6 Intersectorial collaboration
This is an indicator that assesses the frequency of collaboration of doctoral students with partners outside their university. Given that doctoral students are necessarily affiliated with a university, we consider a paper to be the result of an intersectorial collaboration when 1) the paper is authored by a doctoral student and 2) at least one
A non-university organization (industry, hospital, federal government, provincial government, others sector) appears on the paper. The same calculation is used for compiling university researchers and professors’ intersectorial collaboration rate. One of the limitations of this indicator is that the attribution of a sector is limited to Canadian institutions. Thus, a paper written in collaboration with a foreign industry will be considered as the result of international collaboration, but not of intersectorial collaboration.

### 3.6.7 Student-student or faculty-student collaboration

Articles authored by two doctoral students from Quebec will be considered as the result of student-student collaboration. Similarly, papers authored by a professor from Quebec and a student from the same province will be considered as the result of a faculty-student collaboration. This means that articles authored with doctoral students from outside the province, as well as with faculty members from non-Quebec universities, are not here considered as student-student or faculty-student collaborations.

### 3.6.8 Network analysis

In order to visualize the collaborative ties between doctoral students and Faculty members, a network analysis was performed using UCINET (Borgatti, Everett and Freeman, 2002) and Netdraw (Borgatti, 2002) softwares. These softwares allow the creation of 2-dimensional networks of co-authored papers. The size of the edges (lines) between each of the nodes is a function of the number of co-authored papers between the two entities. These networks of collaboration are, thus, based on the collaboration matrix between university researchers’ and students. In most of the
figures, a threshold in terms of numbers of papers written in collaboration is fixed in
order for the network to be clearer.

3.7. Statistical tests of significance for citation measures
Statistical tests were performed using Statistical Package for the Social Sciences,
version 12.0 (SPSS, 2004) to determine if the observed differences between the
scientific impact of publications to which doctoral students’ contributed and 1) those
to which faculty members’ contributed and 2) other Quebec papers were significant.
These tests were performed because of the non-parametric nature of citation
measures (Seglen, 1992), a minority of papers accounting for the majority of
citations. These ‘highly-cited’ papers can have a strong influence on the average
citation measures obtained—one paper cited 100 times weights the same in the
‘average’ as 100 papers cited once—and, hence, create observed difference between
two values that are in fact only determined by the presence in one of the datasets of a
few of these highly cited papers.

A non-parametric statistical test, the Mann-Whitney U (Mann and Whitney, 1947)
was thus used. The Mann-Whitney U is basically a two samples—or populations, as
it is the case in this thesis—t-test that is performed after replacing the actual values
of the two datasets by a combined ranking32. In other words, it is a t-test calculated
on ranks. Three levels of significance are presented: p< 0.001 (highly
significant=***), p< 0.01 (very significant=**), p< 0.05 (significant=*)..

32 For more details, see: http://en.wikipedia.org/wiki/Mann-Whitney_U
4. Socialization and integration of PhD students into research

This chapter presents the bibliometric results obtained on Quebec doctoral students’ socialization to research. As mentioned in the introduction of this thesis, scientific publishing is a central aspect of performing research. It could be considered as the last step of a research project, where the research results are diffused to the community. Indeed, knowledge newly created by researchers—PhD students or professors—has to be validated and diffused, and it is through the process of scientific publication that both these steps are performed. The new knowledge is embedded into a scientific document (article, conference proceeding, book, etc), whose diffusion into print (or online) depends on the outcome of its evaluation made by peers. Although doctoral students have to write a thesis and have it evaluated by a committee of peers, this is done in a controlled environment. It is also likely to be the last thesis of the doctoral student (unless the student decides to enrol in a second PhD program). On the other hand, if the doctoral student wishes to pursue a career in research, writing scientific papers will be an important part of his/her life.

For doctoral students, contributing to research and, ultimately, to scientific papers is a new enterprise, to which they are being socialized during their doctoral studies. Through the bibliometric analysis of doctoral students’ publishing activities, this chapter analyses three aspects of their socialization to research. Firstly, using the

“With the system functioning as it does, it appears that much physics research is conducted only because there happens to be a postgraduate available who is interested enough to do it.” [Walford, 1983, p. 252]
percentage of doctoral students who published at least one paper during the course of their PhD program, it measures the extent to which they are, as a social group, socialized to research. Secondly, the context of this socialization is studied, using author and institutional address characteristics found on papers. Finally, the effect of the participation in peer-reviewed papers on their subsequent research careers is measured for the specific cohorts, using the pre- and post-graduation publication patterns.

4.1. The extent of doctoral students socialization to research

The percentage of doctoral students who participate in peer-reviewed publication provides an indicator of the extent to which, globally, they are integrated and socialized to research. The higher this percentage is, the more we can consider doctoral students of a given discipline to have experienced a certain kind of socialization to research. Figure 13 presents, by first year enrolled and discipline of doctoral program, the percentage of doctoral students who have published at least one paper over the course of their PhD. Unsurprisingly, basic medical sciences is the discipline where the largest proportion of doctoral students are socialized to research during their PhD via the publication of peer-reviewed papers, with a peak at 79% in 2001. It is followed by the sciences (peak at 61%), health sciences (58%) and engineering (46%) disciplines which are all above the average.

---

33 Two typologies of disciplines are used in this thesis: the discipline of the journal in which the articles are published (presented in section 4.1.1, tables 2-3) and the discipline of the doctoral program in which doctoral students are enrolled (section 4.2, table 8). The former classification is used when the characteristics of doctoral students’ papers are compared with those of other Quebec papers, while the latter is used when doctoral students’ characteristics are compared for different years active in the doctoral program or gender, or when they are compared with those of established Quebec researchers (faculty).
On the other hand, only a minority of doctoral students had the experience of publishing at least one journal article in the disciplines of social sciences (with a peak at 20%), business & management (14%), non-health professions (12%), education (8%) and humanities (6%). Although the disciplines used are not exactly the same, the disciplinary distribution of our data follows a trend similar to that of Nettles and Millett (2006). Indeed, they show that doctoral students in the disciplines of engineering (47%) and science (44%) had the higher proportion of output during their PhD, followed by social sciences (22%), humanities (19%) and education (15%). It is quite interesting to note that, even though Nettles and Millett’s (2006) data are based on a survey and, hence, theoretically include all journal articles authored by doctoral students, our numbers—based on the number of papers indexed in the WOS and, thus, a subset of PhD students’ research output—are higher for disciplines of natural and health sciences. These data also follows a trend similar to the sample of disciplines studied by Lee (2000). This shows that Quebec doctoral students are more active in publishing papers than were doctoral students in the United States in 1996, the year Nettles and Millett’s collected their data. The drop in the percentage of doctoral students having published for more recent years also suggest that their participation in papers does not occur during the first years of their doctorate, but rather once they have spent a certain number of years in the program.
Figure 13. Percentage of doctoral students having published at least one paper, by first year enrolled in a doctoral program (cohort) and discipline of the doctoral program, 2000-2007
Several factors can explain the disciplinary differences observed in Figure 13. First and foremost, participation in peer-reviewed papers is increasingly considered, in the disciplines of the medical and natural sciences, as one of the requirements for the completion of the doctoral degree—since theses in these disciplines often taken the form of a series of articles (Breimer, 2010; Holdaway, 1994). Although there is no large scale data currently available on the extent of these practices, it seems that from the literature that the ‘standard’ monograph thesis has been replaced, in most of these scientific disciplines, by a series of articles which have to be published in peer-reviewed journals (thesis as a series articles or manuscript-based thesis)\(^{34}\). It is thus normal that in these disciplines, doctoral students publish scientific papers, since it is now often considered to be one of the requirements for degree completion because of the form of the thesis. On the other hand, in the disciplines of the social sciences and of the humanities, the format of the thesis is still that of the monograph—which is similar to the book format often used by researchers for disseminating research. As a consequence, the writing of articles might be considered less important in their research training.

Another aspect is the structure of the student-supervisor relationship. In the social sciences and the humanities, students typically work from home, often with less interaction with their supervisors than in the natural sciences and medicine (Dellamont, Atkinson and Parry, 1997; Pole et al., 1997; Ridding, 1996). They also tend to be only remotely involved in their advisors’ research, as most professors do

not have any team of their own and prefer to perform research alone (Larivière, 2007). In that sense, it is normal that the highest proportion, among the SSH disciplines, of doctoral students contributing to papers is observed in the discipline of social sciences, as the use of quantitative methods allows more opportunities for labour division and, hence, for collaboration (Moody, 2004).

On the other hand, in the natural and medical sciences, students typically go to the lab every day, and get to work, in addition to their own doctoral research, on other researchers’ research projects (other students, post-doctoral fellows, professors, etc.). Moreover, their research is often one component of a bigger project on which other researchers in the group work, and students typically have difficulty distinguishing their own work from that of their supervisors (Pole et al., 1997). Hence, if one uses the participation in peer-reviewed papers as an indicator, doctoral students in the natural sciences and medicine are more socialized and integrated to research—both their own as well as that of other researchers in the lab—than their colleagues of the social sciences and humanities.

Along the same lines, we have showed in section 3.1 that the contribution needed to sign a scientific paper varies greatly by discipline (Pontille, 2004). In natural and medical sciences’ research teams, senior co-authors habitually grant authorship to junior staff involved in various parts of the research. On the other hand, in disciplines of the social sciences and humanities, because of the historical authorial figure, established researchers are less disposed to grant co-authorship to subordinates—even when they participated in the research (Pontille, 2004). These
different authorship attribution methods might explain, at least in part, the differences observed.

Finally, another factor that might explain the differences is the fact that the WOS—and in fact any existing bibliometric database to this date—does not systematically index books published, nor has a complete coverage of the local literature published in these disciplines. Hence, the lower proportion of doctoral students in SSH active in research might be a reflection of the fact that their scientific production is underestimated, as they are more likely to publish books (or book chapters) or to publish in more local journals.

How does this percentage compare with that of established researchers? Figure 14 compares the percentage of doctoral students (2001 cohort and all cohorts) having published at least one paper over the 2000-2007 period with that of Quebec university researchers. Naturally, a higher proportion of university professors and researchers are active in research than doctoral students, with the exception of basic medical sciences where a larger proportion of doctoral students enrolled in a doctoral program in 2001 than faculty members have published a paper.\(^{35}\)

More specifically, for all disciplines combined, 27% of all doctoral students, 34% of those who enrolled in 2001 and 59% of university researchers have published a paper between 2000 and 2007. The higher proportion observed for the 2001 cohort is caused by the expected fact that doctoral students are more active after a few years in

\(^{35}\) This is likely due to clinical professors, whose main task is not necessarily to perform research.
the program than during their first years. Globally, this figure shows that disciplines in which university researchers and professors publish papers are also those in which a higher proportion of doctoral students are involved in research (sciences, engineering, basic medical sciences, and health sciences). On the other hand, in the disciplines of education, non-health professions, humanities and business and management, a minority of professors publish papers and hence, an even lower proportion of doctoral students do so.

Figure 14. Percentage of doctoral students and faculty members having published at least one paper, by discipline of the doctoral program, 2000-2007

Socialization generally involves role models (van Maanen and Schein, 1979). In the case of socialization to research, established researchers have an important mentoring role. If in a given discipline, established researchers do not publish articles—either because they are not very active in research or because they write
books or other type of scientific literature—it could be expected that doctoral students will not. On the other hand, if in another discipline most faculty members are publishing articles, chances are that doctoral students also will.

To provide more insight on the relationship between faculty and doctoral students’ publishing activities, we built a scatter plot of the percentage of doctoral students with at least one paper, as a function of the number of faculty members with at least one paper (Figure 15). Each dot represents a different doctoral program specialty. The scatter plot shows a very strong exponential relationship between the two variables ($R^2=0.78$). There is a group of specialties in the lower left hand side of the Figure where the percentage of faculty with at least one paper varies from 12% to 50%, but in which only less than 10% of doctoral students are authors of at least one paper. Another group of specialties, in the upper right hand corner of Figure 15, exhibit an opposite pattern. In these specialties, the percentage of publishing faculty spans between 60% and 80%, while that of doctoral students varies from 20% to almost 70%. Hence, although the proportion of professors having published at least one paper follows a more or less normal distribution, the same data for doctoral students is more skewed: in a majority of specialties a minority of students published papers, while in a minority of specialties, a majority have done the same.

The ratio between the percentage of doctoral students and the percentage of faculty members with at least one paper is highly correlated, using a power law fit, with the percentage of doctoral students with at least one paper (Figure 15, inset). Indeed, in disciplines where only few doctoral students participate in peer-reviewed papers, the
difference between the percentage of students and the percentage of faculty with papers varies quite a bit among disciplines. For example, in law and in philosophy, respectively 2% and 3% of doctoral students have published papers, compared with 28% and 50% for university-based researchers and professors, giving a professor/student publishing ratio of more than 14 (which mean that faculty members are more than 14 times more likely to be the authors of at least a paper than doctoral students). On the other hand, in laboratory medicine and in astronomy and astrophysics, these ratios are only 1.3 and 1.5, respectively, with 85% of university professors’ publishing and 58% and 68% of doctoral students having published in each of these two areas.

Globally, Figure 15 reveals that there are two types of disciplines, which basically follow the usual dichotomy between, on the one hand, the social sciences and humanities and, on the other hand, medicine the natural sciences and engineering. In the former, only a small fraction of students get to publish papers, irrespective of whether professors in the field are active or not in research, while, in the latter, doctoral students’ participation in papers closely follows that of established researchers. This suggests that, in addition to the disciplinary patterns in authorship attribution, collaboration affects doctoral students’ participation in papers, as professors in the SSH might be less likely to include their student colleagues into their own research projects, while researchers in the natural sciences and medicine are more likely to do so. The effect of collaboration between doctoral students and faculty will be explored in the next section.
These different ways of organizing scientific research also reflect on the research funds which are attributed to researchers to, among other things, hire research staff such as doctoral students, post-doctoral fellows or professional associates. Indeed, a larger proportion of researchers in the natural and medical sciences than in the social sciences and humanities hold grants (Larivière, Macaluso, Archambault and Gingras, 2009), and the typical grant is much smaller in the social sciences and humanities than in the natural and medical sciences. On the other hand, the amount (in dollars) of doctoral scholarships given by the research councils is about the same in each of
Thus, as shown by many authors (e.g., Buchmueller, Dominitz and Hansen, 1999; Nettles and Millets, 2006), access to research funds—be it in terms of research assistantships, teaching assistantships or fellowships/scholarships might also explain doctoral students’ involvement in peer-reviewed papers. Although we do not have, in this thesis, any data on students’ own funds (scholarships, etc.) we do have data on faculty research funding received, thanks to the SIRU database (Système d’information sur la recherche universitaire—Information System on University Research) of the MELS.

Figure 16 presents the relationship between the average amount in funding received by professors and the percentage of doctoral students who published at least one paper, by specialty of doctoral program. In order to take into account a time lag between funding and publications, funding is compiled for the 1999-2006 period, while publications are—like in most figures of this thesis—compiled for the 2000-2007 period. Figure 16 provides evidence of a strong correlation between the two variables, which mean that research funds of professors are related to students’ participation in papers. This is not surprising, as research funds received by faculty

---

36 Two different doctoral scholarships are available at the federal level via NSERC (Natural Sciences and Engineering Research Council of Canada), SSHRC (Social Sciences and Humanities Research Council of Canada) and CIHR (Canadian Institutes of Health Research): the Canada Graduate Scholarships – Doctoral (35,000$ a year for a maximum of three years) and Postgraduate Scholarships – Doctoral (20,000 a year for a maximum of four years). NSERC was funding about 2,500 doctoral students in Canada in 2008-2009 via these programs, while SSHRC funded slightly more (about 2,700). On the other hand, CIHR funds fewer students directly: 517 via the Canada Graduate Scholarships – Doctoral and 404 via other doctoral programs. This suggests that the funding of students in health disciplines more often comes from advisors’ research funds or from other sources (foundations, pharmaceutical companies, etc.) rather than from the ‘excellence’ scholarships. In Quebec, the doctoral scholarships given by the FQRNT, FQRSC and FRSQ are 20,000 annually for a maximum of three years.

37 More detail on these data can be found in Larivière et al. (2009).
members are often used to give scholarships or to hire students to work on their research projects which, in turn, increase students’ participation to peer-reviewed papers as well as faculty members’ probability of being funded again. This ‘external’ funding might even have a stronger global effect on students’ participation to papers than the ‘excellence’ scholarships given by research councils, since the latter type of funding might increase students’ independence from their advisors and, thus, reduce the probability that they participate in their advisors’ projects. Further research is needed here to assess these potential effects.

**Figure 16. Relationship between the average amount in funding received by professors and the percentage of doctoral students who published at least one paper, by specialty of doctoral programs, 2000-2007 (1999-2006 for funding)**

Several studies have examined gender differences in productivity (see, among others, Fox, 2005; Prpić 2002; Xie and Shauman 1998 and 2003; Zuckerman, 1991). These studies all show that women were, on average, publishing fewer papers than their
male counterparts. However, no study has yet analyzed gender differences in doctoral students’ participation in peer-reviewed papers. Figure 17A presents the percentage of doctoral students having published at least one paper, by gender, as well as the percentage of women PhD students, by discipline of the doctoral program. It shows that, for all disciplines combined, a higher proportion of male than female students authored at least one paper (34% vs. 27.5%).

However, this difference is not observed in all disciplines. Male PhD students tend to be more involved in research than their female colleagues in basic medical sciences and in health sciences, as well as, to a lesser extent, in business and management, in the humanities and in engineering. On the other hand, a higher proportion of female students are publishing in education, non-health professions and, albeit only slightly, in social sciences and natural sciences. That being said, breaking down the data at the level of specialties of departments (Figure 17B) provides evidence that in most specialties (28 out of 42), male doctoral students are more active in research than their female colleagues. On the whole, medical disciplines—where most of the research is performed in laboratories—seem to favor male students, as a higher proportion of them are integrated to research. The opposite is true for more individualistic disciplines such as the social sciences, non-health professions and education.
Figure 17. A) Distribution of the percentage of doctoral student having published at least one paper, by discipline of the doctoral program and gender, 2000-2007. B) Distribution of the difference between male and female PhD students participation in peer reviewed papers, by specialty of the department.

A) Percentage of doctoral students

- Men
- Women
- % Women in the discipline

B) Difference men/women

Specialty of the department
4.2. The context of socialization

Using bibliometric methods, one can produce an indicator of the context in which doctoral students participate in peer review papers and are socialized and integrated to research. Two type of information appearing in the by-line of scientific papers and indexed in Thomson Reuters’ database are used to provide contextual information: the co-authors and institutional addresses.

Figure 18 presents the percentage of doctoral students’ papers that are co-authored compared to other Quebec papers (excluding papers to which doctoral students contributed). It shows that, in each discipline, PhD students’ papers are more likely to be authored with other researchers. Although in most disciplines, the quasi-totality of the papers is co-authored (see, among others, Cronin, 2005; Larivière, Gingras and Archambault, 2006), for doctoral students, the proportion is even higher, with about 99% of papers having more than one author in biology, biomedical research, chemistry, clinical medicine, earth and space, engineering, physics and psychology. Unsurprisingly, only the humanities display a different pattern: only 10% of the papers are co-authored, with only a very small difference between doctoral students’ papers and other Quebec papers. This strongly suggests that, in their apprenticeship of research, doctoral students benefit from the help of other researchers who are likely more experienced and established. Although we do not have any information on the link between the student and the other authors, it could be expected, given the qualitative research on doctoral students’ socialisation to research, that these co-authors are, more often than not, their supervisors/mentors (Pole et al., 1997), although other social groups—such as postdoctoral fellows—might be among these
co-authors and are also likely to play an important role in the socialization to research of doctoral students (Delamont, Atkinson and Parry (1997).

Similarly, Figure 19 provides data on the average number of authors of doctoral students’ papers and of other Quebec papers. As one might expect from the Figure previously discussed, the mean number of authors is higher for papers to which doctoral students have contributed than for other Quebec papers, as the PhD student was an ‘extra’ author added. More specifically, for all disciplines combined, papers to which doctoral students contributed have twice as many authors than Quebec papers to which they did not contribute (18.9 vs. 9.8). Two disciplines exhibit a different pattern: the humanities, where the number of authors is the same for PhD students’ papers and for other Quebec papers and physics, where the number of co-authors of students’ papers is twice as high as that of other Quebec papers (161 vs. 78). In the case of physics, this is likely due to the very high number of co-authors of high energy physics papers, as well as the higher probability, given authorship attribution practices of that specialty (Pontille, 2004) that doctoral students get the opportunity to sign papers.
Figure 18. Percentage of papers with more than one author (co-authorship) of PhD students’ papers and of other Quebec papers, by discipline of journal, 2000-2007

Figure 19. Average number of authors of PhD students’ papers and of other Quebec papers, by discipline of the journal, 2000-2007
These two figures also show that, in all disciplines but the arts and humanities—where the often non-empirical nature of research reduces the need for labour division and, hence, for collaboration (Moody, 2004)—doctoral students’ socialization to research is rarely a one-to-one (student and advisor) relationship, but is often done within the context of a research team (Gemme and Gingras, 2008).

In order to gain more insight into the status of the co-authors of doctoral students, we compiled the percentage of doctoral students’ papers authored with faculty members as well as with other doctoral students from Quebec universities (Figure 20). It reveals that in all disciplines but the humanities—where only a small proportion of papers are co-authored—more than 90% of PhD students’ papers are authored with faculty members. Other doctoral students are also co-authors of a certain percentage of papers, which varies between 6% in the humanities and 34% in education. These numbers strongly demonstrate the importance of senior and established researchers in doctoral students’ apprenticeship of research. However, these senior researchers are not the only social group involved in the process, as a significant part of PhD students’ papers are also written in collaboration with other student colleagues. Unfortunately, we do not have data on post-doctoral fellows—their status is unclear and varies among universities—but one could expect them to account for a significant proportion of the collaborators, at least in the natural and medical sciences.
Figure 20. Percentage of doctoral students’ papers written with faculty members or with other doctoral students from a Quebec university

Figure 21 looks the question the other way around and presents the percentage of faculty papers that are authored with doctoral students and with other faculty members. As one could expect, the percentage of faculty members’ papers authored with doctoral students is lower than the percentage of PhD students’ papers authored with faculty members. Still, in disciplines of the natural sciences and medicine, more than 30% of faculty papers are authored with doctoral students. This percentage is slightly lower in the social sciences (27.4%) and education (25.5%), below 20% in non-health professions and business and management and below 10% in the humanities. This globally points up the relative importance of doctoral students among established researchers’ collaborators: in disciplines where data collection is important, doctoral students account for an important part of established researchers’ collaborators; in disciplines that are less data intensive, such as the humanities, this percentage is lower.
In all disciplines, however, collaboration with other faculty members is more important than collaboration with doctoral students. In basic medical sciences and health sciences, this suggests that collaboration between different labs is quite frequent, and that these collaborations do not necessarily involve doctoral students. Along the same line, in education, social sciences and business and management, faculty members tend to work more often with other established researchers. On the other hand, in sciences and in engineering, faculty collaborate as much with doctoral students as with other faculty members, which suggest that there is less ‘inter-laboratory’ collaboration in these disciplines.

**Figure 21. Percentage of faculty papers authored with doctoral students from a Quebec university**

In order to obtain a better understanding of the number of individuals involved in faculty/students collaborations, we compiled the average number of distinct faculty collaborators per PhD student, as well as the average number of distinct PhD
student collaborator per faculty (Figure 22). We see that in all disciplines but sciences
and engineering, PhD students are collaborating with a higher number of faculty
members than the latter do with students. Analyzed in parallel with Figure 23 (A, B,
C, D)—which presents doctoral students and faculty members’ co-authorship
network in four disciplines—this shows that doctoral students collaborate with
several established researchers and thus might benefit from a wide variety of
mentors.

In sciences and engineering, where most established researchers typically have their
laboratory, we see that there are more students gravitating around faculty (intra-lab
collaboration) than the inter-laboratory or inter-faculty collaboration. This is
exemplified by Figure 23B, which presents the faculty/PhD student collaboration
network in physics and astronomy.
Figure 22. Average number of collaborating Quebec faculty in PhD students’ papers and average number of PhD students in Quebec faculty papers, by discipline of the department, 2000-2007

Figure 23 also exhibits the different collaboration densities: in laboratory medicine, even with a threshold of more than 2 papers, the network is very dense, with both faculty and PhD students having a lot of collaborators. This is also a reflection of the size of the domain (N=2,410 doctoral students over the 2000-2007 period). On the other hand, the network in physics and astronomy (Figure 23B, 577 doctoral students over the period) is less dense, and often structured with one established researcher around which several doctoral students gravitate; a good visual representation of the lab structure of scientific discipline, one professor is the head of the laboratory and has several masters’ and doctoral students, post-doctoral fellows as well as other professional and technical staff working with him.
Figure 23C and D presents, with no threshold, the networks for psychology (3,071 doctoral students over the period) and education (1,536 doctoral students), respectively. One can readily see that the network is denser in psychology than in education, with PhD students collaborating with several senior researchers and senior researchers collaborating with several doctoral students. In education, the trend is different, however: we clearly see that doctoral students collaborate with several professors who, in turn, collaborate with a small number of students. This corroborates Figure 22, which showed that the number of collaborating professor per student was more than three times as high as the number of collaborating doctoral student per professor.
Science is a collective endeavour, and collaborators are increasingly found in foreign countries. Indeed, researchers throughout the world are increasingly collaborating with foreign partners (see among others Georghiou, 1998; Glänzel, 2001; Leydesdorff and Wagner, 2008) and Quebec researchers are no exception (Larivière, 2007). Figure 24 presents the percentage of doctoral students’ papers and of other Quebec papers that are authored in collaboration with foreign colleagues, by discipline of the journal in which the papers are published. Although all collaboration
measures presented so far in this section provided evidence that doctoral students’ papers were more likely to be the result of collaboration, Figure 24 provides a distinct picture. Indeed, in all disciplines but the arts and the humanities, a significantly lower proportion of doctoral students’ papers than of other Quebec papers have international co-authors. Hence, although doctoral students’ integration to research is generally made in bigger teams, these teams are less likely to include international collaborators. One could think that, given their apprenticeship status in research, doctoral students might not be attributed international projects by their supervisors in natural sciences and medicine, or don’t have yet international contacts to publish with in the SSH.

**Figure 24. Percentage of international collaboration of doctoral students’ paper and of other Quebec papers, by discipline of the journal, 2000-2007**
University-industry relationships have generated a lot of literature since the seminal book of Gibbons et al. (1994). Several studies have made evident the increase of such links in publications (Leydesdorff and Sun, 2009) and its effect on scientific impact (Lee, 1996, Godin and Gingras, 2000b; Lee and Bozeman, 2005; Lebeau et al., 2008); others have analyzed how doctoral students are involved in such collaborations (Gemme, 2009) or have worked on university-industry funding programs (Vécrin, 2003).

Figure 25 presents the proportion of faculty and doctoral students’ papers authored with researchers from outside academia and university hospitals, that is, researchers from industries and governments, as well as from other non-university organizations (non-for profit organizations, cities, etc.). Although the global difference between PhDs and faculty is negligible for all disciplines combined (9.8 vs. 9.3), it is important to note that PhD students’ collaborations with other sectors are behind those of faculty only in health sciences. In each of the other disciplines, doctoral students are more likely to co-author papers with researchers from outside the university sector. Hence, despite the small difference observed at the level of all disciplines, the fact that a difference is observed in almost all disciplines suggests that partners outside academia may be important actors of doctoral students’ socialization to research, which could be explained by an increase in the percentage of doctoral students who perform their doctoral research in a non-university setting (Gemme, 2009, Gemme and Gingras, 2008; Thune, 2009). The very high difference observed in business and management, education and non-health professions suggests that such collaborations are also very important in non-scientific disciplines, although these are more often
with governmental organizations rather than industries. One cannot also exclude the fact that doctoral students in these disciplines are often older and might already have jobs in these non-academic organizations and add them as institutions of affiliation on the papers they author during their doctorate. These data are consistent with those of Gemme and Gingras (2008), who also found that students in the social sciences and humanities were more likely to have links with non-university organizations than their colleagues of the natural and medical disciplines.

**Figure 25. Percentage of papers authored with researchers from sectors other than universities and hospitals, 2000-2007**

Previous figures showed that the vast majority of doctoral students’ papers were co-authored. But where do doctoral students stand in this author list? As mentioned previously in this thesis, in all disciplines but mathematics and high energy physics, the first author is generally considered as the main contributor to the paper. Hence, the closer doctoral students are to the first authorship position, the more we can
consider them as the main individuals responsible for the research project. Figure 26 presents the percentage of doctoral students’ papers for which they are the first authors by discipline of the journal. In all disciplines but health, physics and professional fields, doctoral students are first authors of more than half of the papers to which they contributed. Unsurprisingly, in the arts and in the humanities—where a higher proportion of papers are sole authored—, doctoral students are the first authors of more than 70% and 90% of the papers the author, respectively. On the whole, this strongly proves that, in the majority of cases, doctoral students have the primary responsibility of the papers to which they contributed.

**Figure 26. Percentage of Quebec’s doctoral students’ papers that are first authored, by discipline of the journal, 2000-2007**

[Bar chart showing the percentage of first-authored papers by discipline.]
Given that the average number of authors differs greatly among disciplines and, hence, to make data more comparable across disciplines, we have divided, for each paper authored by doctoral students, the students’ order in the author list by the total number of authors of the paper\textsuperscript{38}. When the normalized author order is close to 1, the doctoral students (or faculty) are, on average, near the last author position; when they are close to 0, they are near the first author position. For example, the first author of a paper with 10 authors obtains a normalized author order of 0.1, and the ninth author of the same paper will obtain a normalized author order of 0.9.

Figure 27 compares this index by discipline of the doctoral program for A) all papers authored by either PhD students or faculty as well as B) for papers in which there is a faculty/doctoral student collaboration. In each of the disciplines, when all of their papers are considered, we clearly see that doctoral students are closer to the first authorship position than faculty members. This confirms the findings presented in the preceding figure, but also provides additional empirical evidence of the well known fact that—at least in laboratory-based disciplines—professors generally sign as last author the papers coming out of their laboratories (Boyack, 2009).

When only the subset of papers co-authored by doctoral students and faculty are considered, this tendency remains the same in practically all disciplines: faculty are closer to the last author position and doctoral students remain closer to the first. However, the trend differs in business and administration and in the humanities: in these disciplines, senior researchers are more likely to be the first authors in case of

\textsuperscript{38} This indicator was previously used in Gingras, Larivi\`ere, Macaluso and Robitaille (2008).
collaboration with doctoral students. This strongly suggest that there is little *noblesse oblige*—the act of granting more visibility (Zuckerman, 1977)—in these disciplines. This might also be a reflection of the historical figure of the sole author (Pontille, 2004) and of a more commanding academic hierarchy, which could prevent established researchers from granting first authorship to younger colleagues. That being said, one cannot exclude the possibility that, in these two disciplines, doctoral students have a smaller contribution to the papers they author than that of their senior collaborators and, hence, generally sign the papers as second or third author. This might especially be the case in the humanities, where the theoretical aspect of research is often more important than the empirical one and where senior researchers, because of their greater knowledge of past literature, might provide more input in papers than their student co-authors. Further qualitative research could shed more light on these aspects.
Figure 27. Normalized author order of A) doctoral students and faculty papers and B) for papers co-authored by at least one doctoral student and one faculty member, by discipline of the doctoral program, 2000-2007.
As mentioned previously, mentoring relationships are an important component of doctoral students’ socialization to research (Nettles and Millets, 2006). In this respect, having mentors of the same gender might affect doctoral students’ participation in papers—especially for women, who are generally underrepresented in academia (Neumark and Gardecki, 1998; Xie and Shauman, 2003). Table 13 presents, for the subset of papers authored by Quebec doctoral students and faculty members, the percentage of collaboration of each gender, by gender of the collaborating faculty. Given the general underrepresentation of women in faculty positions, the percentage of doctoral students’ collaboration with women faculty is lower than with male faculty. However, by comparing the difference of men and women doctoral students’ collaborations with men and women faculty, there may be a gender affinity in the choice of collaborators.

As one could expect, there is indeed a strong tendency by doctoral students to collaborate with established researchers of the same gender. More specifically, for all disciplines combined, while 86.9% of male students’ collaborations with faculty are with men, this percentage drops to 78.5% for women students. To the opposite, 13.1% of male students’ collaborations are with women faculty, compared to 21.5% for female doctoral students. The only discipline where no such difference is observed is education, which is also the discipline with the highest proportion of women doctoral students (Figure 17) and the second highest proportion of women faculty (CREPUQ, 2009).
<table>
<thead>
<tr>
<th>Discipline and gender of PhD students</th>
<th>Gender of faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Basic Medical Sciences</td>
<td>83,9%</td>
</tr>
<tr>
<td>Male</td>
<td>86,0%</td>
</tr>
<tr>
<td>Female</td>
<td>81,6%</td>
</tr>
<tr>
<td>Business &amp; Management</td>
<td>74,8%</td>
</tr>
<tr>
<td>Male</td>
<td>84,8%</td>
</tr>
<tr>
<td>Female</td>
<td>50,0%</td>
</tr>
<tr>
<td>Education</td>
<td>70,2%</td>
</tr>
<tr>
<td>Male</td>
<td>70,8%</td>
</tr>
<tr>
<td>Female</td>
<td>70,1%</td>
</tr>
<tr>
<td>Engineering</td>
<td>89,9%</td>
</tr>
<tr>
<td>Male</td>
<td>91,0%</td>
</tr>
<tr>
<td>Female</td>
<td>85,1%</td>
</tr>
<tr>
<td>Health Sciences</td>
<td>69,0%</td>
</tr>
<tr>
<td>Male</td>
<td>79,0%</td>
</tr>
<tr>
<td>Female</td>
<td>59,7%</td>
</tr>
<tr>
<td>Humanities</td>
<td>72,0%</td>
</tr>
<tr>
<td>Male</td>
<td>81,3%</td>
</tr>
<tr>
<td>Female</td>
<td>38,9%</td>
</tr>
<tr>
<td>Non-Health Professional</td>
<td>70,7%</td>
</tr>
<tr>
<td>Male</td>
<td>76,8%</td>
</tr>
<tr>
<td>Female</td>
<td>66,0%</td>
</tr>
<tr>
<td>Sciences</td>
<td>86,5%</td>
</tr>
<tr>
<td>Male</td>
<td>89,1%</td>
</tr>
<tr>
<td>Female</td>
<td>82,6%</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>72,2%</td>
</tr>
<tr>
<td>Male</td>
<td>78,9%</td>
</tr>
<tr>
<td>Female</td>
<td>66,7%</td>
</tr>
<tr>
<td>All Disciplines</td>
<td>83,3%</td>
</tr>
<tr>
<td>Male</td>
<td>86,9%</td>
</tr>
<tr>
<td>Female</td>
<td>78,5%</td>
</tr>
</tbody>
</table>
These data indicate strong gender effects in socialization, as women students are more likely to collaborate with women, and men with men. These data are consistent with Long (1990), who established, for a subset of biochemistry researchers, that women were more likely to have women mentors. Given that women tend to publish less than men (Fox, 2005; Prpić 2002; Xie and Shauman 1998 and 2003; Zuckerman, 1991), these gender affinities—which are likely a reflection of doctoral students’ choice of advisors—might explain why a smaller proportion of women than men have published at least a paper during their doctoral studies (Figure 17).

4.3. The impact of socialization on their subsequent research careers

The act of publishing is an indisputable effect of socialization to research. Several authors have analyzed the impact of socialization to research on students’ ulterior research careers. More specifically, authors such as Turner and Thompson (1993) and Gardner (2007) have positively linked socialization to research with degree-completion. Nettles and Millets (2006) established that research productivity was positively linked with degree completion and time spent in the program, i.e., students with a publication record did not spend more time in the program. Similarly, Seagram, Gould and Pyke (1998) observed that collaborating with doctoral advisees on conference papers was one of the factors reducing time to completion.

Figure 28 presents the number of papers by doctoral students of the 2000 and 2001 cohorts, for doctoral students who completed their doctorate as well as those who had not completed it as of the end of 2007. It clearly demonstrates that, in each of
the disciplines, those who had completed their doctorate published a higher number of papers than those who had not completed the program. These data provide strong evidence of the relationship between participation in the research activity and degree completion. Indeed, an important aspect of the doctorate is to contribute to the advancement of scholarly knowledge in a discipline. It is thus normal that, by publishing papers—which are contributions to knowledge—doctoral students increase their chances of completing their doctoral degrees. On the other hand, the average impact per paper is not significantly different for the two groups of students (not presented in the Figure). That being said one cannot exclude the possibility of a strong ‘selection’ bias, where those who have published and graduated were actually the ‘best’ PhD students.

Although one could argue that, in the natural sciences and medicine, these differences are caused by doctoral theses that consist of a series of published articles—which now seems to be the standard form of the thesis in these disciplines (Breimer, 2010; Holdaway, 1994)—these practices seldom exist in the social sciences and humanities and, hence, cannot explain the differences observed in these disciplines, which are as high as those observed in the natural sciences and medicine. On the whole, these data shows that being integrated into research is strongly linked with doctoral students’ degree completion.

39 By definition, this excludes doctoral students who have not published any paper.
To measure the relationship between pre-graduation integration into research on post-graduation research activity, we compiled the average number of post-graduation papers as a function of classes [0, 1, 2, 3, 4, 5 or more] of pre-graduation papers (Figure 29). In order to have enough numbers of pre- and post-graduation papers, only the subset of doctoral students who had graduated in 2003 or 2004 are considered.

Figure 29 clearly shows that there is a positive relationship between the two variables; those who publish more during their doctorate are more likely to publish more afterwards. Although the trends are less clear in disciplines of the social sciences.

---

40 One of the limitations of Figure 29 is that it only includes post-graduation papers having at least one Quebec address, as this was one of the matching criteria for assigning the student’s papers. Hence, papers authored during a post-doctoral fellowship abroad are not included, except when they are written with collaborators from Quebec. This thus reduces the...
sciences and humanities, we observe that in all disciplines, doctoral students who had not authored any paper during their studies obtain lower publication rates after graduation. This is likely an effect of socialization and integration into research: students who are more involved in research during their doctorate are socialized to the publication *habitus* (Bourdieu, 1980; 2001) and keep this *habitus* after graduation, as they themselves become members of the scientific community.
Figure 29. Relationship between pre-graduation productivity and post-graduation productivity, for the subset of doctoral students who graduated in 2003-2004 (N=2,319)
This chapter provided empirical data on the extent of doctoral students’ socialization and integration into research as expressed through their publishing activities, on its context and on its impact on students’ ulterior research careers. It revealed that about a third of doctoral students had an academic publication experience during their studies; a percentage that is similar, albeit smaller, to that obtained by Gemme and Gingras (2008) for a sample of doctoral students in Quebec. The fact that Gemme and Gingras’ data are drawn from a survey—which, by definition, includes all the articles published by students—certainly explains the difference. Given also that a third of Quebec’s PhD graduates are integrated into faculty (Conseil Supérieur de l’Éducation, 2003) it is reasonable to think that there is an overlap between students who publish papers during their studies and those who become faculty members.

The proportion of doctoral students who participated in papers is highly linked to the discipline in which they are active. Indeed, while about four-fifths of doctoral students in basic medical sciences participated in papers, this percentage is below 10% in education and the humanities. Several factors explain these differences, among which the article-based structure of thesis of the natural and medical sciences vs. the monograph thesis of the social sciences and humanities might have an effect, to which we might add the difference in the structure of teams and level of integration of doctoral students into research.

These differences in the integration of doctoral students into research seem to be influenced by collaborative practices. Indeed, in disciplines where collaboration is the norm, senior researchers are more likely to grant co-authorship to their doctoral
students and hence, to include them into their research teams. On the other hand, in disciplines where collaborations are less frequent, a smaller proportion of doctoral students have authored papers during their doctoral studies. The positive effect of collaboration on students’ participation to papers was also observed by Louis et al. (2007).

Analysis of the context of this socialization uncovered that collaboration is the predominant mode of publication: the quasi-totality of doctoral students’ papers is authored in collaboration and in about 90% of the cases with faculty members. They are nonetheless, in a majority of cases, the first authors of the papers they contribute to. On the other hand, a smaller proportion of doctoral students’ papers than for other Quebec papers are authored with foreign partners, which suggest that doctoral students’ socialization is made within projects of a ‘controllable’ size. Gender is also an issue, as female students are more likely to collaborate with female faculty, and male students more likely to collaborate with senior researchers of the same gender.

Finally, this chapter provided clear evidence of the positive link between integration to research, and degree completion and post-graduation research activity. Indeed, doctoral students who have completed their degree have published, on average, three times more papers than students of the same cohorts who have not graduated. Similarly, we see a strong relationship between pre-graduation and post-graduation research productivity, and this relationship is visible in all disciplines.
5. PhD students within Quebec research system

This chapter provides empirical data on doctoral students’ importance in the Quebec research system. It shows the proportion of all papers authored by Quebec researchers to which doctoral students contributed, as well as the comparative citation rates of doctoral students’ papers and of other Quebec papers. Before delving into these results, the first section of this chapter provides contextual data on Quebec’s research system.

5.1. Quebec’s research system

This section presents general characteristics of Quebec’s research system. As mentioned in the introduction of the thesis, Quebec’s research system is defined as the sum of all scholarly research conducted in the province, in all disciplines of the natural or medical sciences as well as the social sciences, the arts or the humanities. The goal of this section is twofold. First, its aim is to provide an overview of the research system in which the population of PhD students evolves. Secondly, it provides evidence that this system is not so different from that of other industrialized countries and that, although this thesis is a large scale study of a the complete population of doctoral students of a Province (Quebec), there are no

---

41 For more empirical data on Quebec and Canada’s research systems, see, among others, Institut de la statistique du Québec (2007; 2008 and 2009), Godin, Doré and Larivière (2002) or Statistics Canada’ periodical service bulletin on science statistics, catalogue 88-001-XIE.
reasons to believe that the trends presented are different from those which could be observed in other Canadian provinces or in any other industrialized countries.

To give an idea of the size of Quebec’s research output, one can compare its absolute number of publications with that of other countries. With 196,814 papers—almost all in English, as it is the case in scholarly communication in general (Ammon, 2001)—indexed in the WOS over the 1980-2007 period, researchers from Quebec have published slightly fewer papers than Brazilian researchers as a whole (211,118), but more than Taiwanese (186,490) and Danish researchers (175,572). Quebec’s scientific production is also considerably higher than that of small countries—often stated as examples of very innovative countries—such as Finland (155,698) and Norway (116,168). More specifically, Quebec researchers increased their scientific production from 3,600 in 1980 to 11,000 in 2007. For Canada as a whole, these numbers increased from 20,000 and 48,000 papers over the same period.

Figure 30 presents Quebec’s papers percentage of Canadian and of World papers. It reveals that, after increasing steadily from 1980 onwards, these percentages peaked in 1995 to 1.2% of the World’s papers and 24% of Canadian papers. Because of budget cuts in the university system, these percentages dropped after 1995 until the first years of this millennium. Since these budget cuts also had an impact on other Canadian provinces, Quebec’s percentage of Canadian papers was less affected that its percentage of World papers, which plunged from 1.2% to less than 1%. As new funding came with the turn of the millennium (Canada Research Chairs program, Canadian Foundation for Innovation, etc., in addition to an increase in budget of the
three Federal research councils), Quebec regained some of the lost ground—albeit not all—and accounted for slightly less than 1.1% of World papers in 2007, and slightly more than 23% of Canadian papers, a number that, roughly matches its share of the Canadian population.

**Figure 30. Quebec’s percentage of Canadian and World papers, 1980-2007**

As in any other research systems, Quebec’s scientific papers are not all authored by university researchers. Although we do not have any comparable data on the share of sectors in other countries, we can nonetheless compare Quebec sectors’ share with that of Canada as a whole. Indeed, while slightly more than 79% of Quebec’s papers have at least one author from universities between 2000 and 2007, this percentage is

---

42 Very few bibliometric organizations have categorized research organizations into their sectors. We expect, however, that the trends are relatively similar for most industrialized countries, with the United States and Japan having a higher percentage of their scientific output authored by researchers from industries, and France, with the CNRS, having a higher proportion of its research performed in governmental laboratories.
of 80.5% for Canada as a whole (Table 14). This Table confirms that the importance of research sectors is roughly the same in Quebec and in Canada. The only exceptions are the Federal government, which is more active in Ontario—because of the many federal laboratories located in Ottawa—and hospitals, which account for a more important share in Quebec than in Canada. The high proportion of papers from hospitals in Quebec is likely a consequence of its specialization in health research.

Table 14. Percentage of Quebec and Canada’s papers, by sector, 2000-2007

<table>
<thead>
<tr>
<th>Sector</th>
<th>Quebec</th>
<th>Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Government</td>
<td>6.3%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Hospital</td>
<td>24.2%</td>
<td>19.5%</td>
</tr>
<tr>
<td>Industry</td>
<td>5.5%</td>
<td>6.1%</td>
</tr>
<tr>
<td>Other</td>
<td>2.0%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Provincial Government</td>
<td>1.8%</td>
<td>2.8%</td>
</tr>
<tr>
<td>University</td>
<td>79.2%</td>
<td>80.5%</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.5%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

Quebec researchers also perform scientific research that is, on average, cited more than the average scientific paper at the World level (significant p < 0.001). Indeed, Figure 31 presents the average of relative citations (see section 4.6.3 for more details in this indicator) of Canadian and Quebec papers since 1980. It establishes that, throughout the period, Canadian and Quebec papers have maintained high citation rates, and that Quebec and Canada’s average impact per paper are quite similar, though Quebec tended to lag behind prior to the mid-1990s.
Where do doctoral students stand in this system? Figure 32 presents the number of doctoral students and of university researchers in each of the disciplines over the 2000-2007 period. One clearly sees that, in all disciplines except basic medical sciences and health, PhD students outnumbered university-based researchers. The positive differential in favor of university researchers in health related disciplines is caused by the fact that MD students are excluded from this Figure (and from the analysis performed in this thesis), as their degree is considered in Canada a first professional degree and, hence, is not part of graduate study. On the whole, this Figure shows that doctoral students are an important part of the academic workforce, at least in terms of raw numbers. The next section will provide empirical data on their contribution to Quebec’s research system in terms of research output.
5.2. Quebec’s papers to which doctoral students contributed

Figure 33 presents unique data on the percentage of Quebec’s university papers to which at least one doctoral student contributed over the 2000-2007 period. For all disciplines combined slightly less than 30% of all Quebec university papers have doctoral students as authors or co-authors. More specifically, in biomedical research, physics and chemistry, doctoral students contribute to more than 35% of all Quebec papers. For other disciplines of the natural sciences and medicine, such as biology, clinical medicine, earth and space, engineering and physics, a percentage between 25% and 30% of all papers are co-authored by doctoral students. The only discipline of the natural sciences with a very low proportion of papers authored doctoral students is mathematics, a finding that could be explained by the lower collaboration rates in this discipline (Newman, 2004).
In disciplines of the social sciences and humanities, a smaller proportion of Quebec university papers benefit from the participation of doctoral students. More specifically, apart from psychology—in which doctoral students contribute to slightly less than a third of Quebec university papers—the percentage of Quebec papers authored by doctoral students varies between from slightly more than 20% in the social sciences related to health, to less than 5% in arts.

As shown in Figure 14 (page 116), doctoral students in disciplines of the social sciences and humanities are less integrated into research. Although the percentage of doctoral students who published papers and the percentage of Quebec papers authored by doctoral students are inevitably related—if no doctoral student publishes at least a paper in a discipline, then doctoral students’ percentage of Quebec papers will necessarily be zero—this relationship is stronger than one might expect. Indeed, in domains where only a small proportion of doctoral students got engaged in publishing a paper, their proportion of Quebec’s scientific output is smaller. In other words, if only a small proportion of students get to participate in peer-reviewed papers, then doctoral students’ overall proportion of Quebec papers will be smaller than if a majority of doctoral students have the opportunity to publish. This is exemplified by the fact that about 80% of doctoral students in basic medical sciences published at least one paper during their doctoral studies and that about 40% of all papers in biomedical research have doctoral students as authors, while only 5-6% of doctoral students in the humanities have published papers and 13% of Quebec’s papers in the same disciplines have doctoral students as authors.
Very few comparable data exists in the literature and, hence, we cannot know if Quebec’s doctoral students account for a larger proportion of the province’s papers than do their colleagues outside Quebec. Indeed, since existing studies on doctoral students’ participation in peer-reviewed papers relied on surveys (Anwar, 2004; Nettles and Millett, 2006) or on a small sample of papers (Blunt and Lee, 1994; Lee, 2000)—compared to the full population of doctoral students used in this thesis—these studies could not provide any information on the proportion of papers, in a given research system, authored by doctoral students. The only dataset that could provide comparable data is that of Liang, Liu and Rousseau (2004), but their paper unfortunately does not provide any information about the percentage of students’ papers in the Chinese research system. For physics, these data are nonetheless consistent with Walford’s (1983) qualitative observations: “[w]ith the system functioning as it does, it appears that much physics research is conducted only because there happens to be a postgraduate available who is interested enough to do it.” (p. 252)

For the medical sciences, the only analogous data that could be found is that of Cursiefen and Altunbas (1998), who examined the research output of students of a German medical faculty, and found that students were the authors of 28% of all papers of the faculty. This percentage is quite slightly smaller than that observed for Quebec doctoral students in the discipline of biomedical research (40%), but only slightly higher than that obtained in clinical medicine (26%). For the social sciences, comparable data can be found in Blunt and Lee (1994) who surveyed authors who published in the Journal of Adult Education/Adult Education Quarterly and found that
students contributed—either as authors or co-authors—to 128 articles published in the journal, accounting for an important proportion (46%) of all papers published in the journal. In the Quebec’s dataset analyzed here, slightly less than 15% of papers of this discipline were authored by doctoral students, which is three times less. Actually, a percentage as high as that found by Blunt and Lee (1994) cannot be found in any of the specialties of the social sciences and humanities, which is likely a reflection of the peculiar nature of the only journal involved in their dataset.

**Figure 33. Percentage of Quebec university papers authored by doctoral students and faculty members, by discipline of the journal, 2000-2007**

Figure 33 also provides related data on Quebec faculty members’ proportion of the provinces’ university output. Unsurprisingly, the quasi-totality (≈90%) of Quebec’s university papers have at least one faculty member as an author, except in the arts, the humanities and in mathematics, where two-thirds to three-quarters of papers are
authored by faculty\textsuperscript{43}. The percentage of papers authored by doctoral students is also below average in these disciplines, which suggests that a variety of actors—such as postdoctoral fellows, independent scholars and masters’ students—are often involved in research in these disciplines. Although the same social groups are as likely to be involved in research in other disciplines, they do so within the framework of collaboration with faculty which, in turn, does not affect the percentage of the Quebec papers authored by faculty.

These data show that doctoral students contribute to a non-negligible proportion of Quebec universities’ papers, a proportion which, however, varies considerably among disciplines. Although collaboration is likely to affect these percentages, one must not forget that the very nature of disciplines also influence the way research is performed and structured. Labour division is the preferred mode of research in the natural and medical sciences because of the complexity of data collection they involve, and this organization of research allows doctoral students to play an active role in the creation of new knowledge\textsuperscript{44} (Shinn, 1988). On the other hand, doctoral students do not necessarily perform the same role in most disciplines of the social sciences and humanities—especially those that are not empirically-driven—where the ‘research group’ structure is not the norm and where data collection, when needed, does not necessarily imply labour division, as the technical skills involved may be considered as less important. Indeed, scholars in the social sciences and humanities do not always need research assistants, since they can often perform, alone or in

\textsuperscript{43} Since the quasi-totality of doctoral students’ papers are written with faculty, data on faculty’s papers include most doctoral students’ papers.

\textsuperscript{44} Legault’s (1993) ethnographic study provided similar results for laboratories of the social and human sciences.
collaboration with other scholars, all the steps needed for their research, from data collection to the writing of the paper. Furthermore, in social sciences laboratories—where doctoral students are involved in research and perform various tasks of data collection and analysis (Legault, 1993)—the act of data collecting does not generally grant authorship to doctoral students (Pontille, 2004).

Figure 34 provides more insight on the effect of collaboration on PhD students’ participation in papers. It uncovers the relationship, at the level of each specialty, between the percentage of papers authored by doctoral students and the percentage of co-authored Quebec papers (Figure 34A), and between the average number of authors per paper (Figure 34B). Both Figures confirm that collaboration is linked to the percentage of Quebec papers authored by doctoral students, although this relationship is not as strong as one might expect. Indeed, when the percentage of Quebec papers that are co-authored is used as a predictor of the percentage of papers authored by doctoral students, a coefficient of determination ($R^2$) of 0.24 is obtained. If the average number of authors per paper is used instead, the $R^2$ drops to 0.11. This points up that, although collaboration is unequivocally linked with students’ proportion of the province’s papers, it only accounts for a small proportion of the variance.

45 In Figure 33B, specialties of General Physics and of Nuclear & Particle Physics are excluded, as they averaged 106 and 177 authors per paper, respectively.
Even though the time span of this study is not very long compared to that of other bibliometric studies—Larivière, Gingras and Archambault (2009) have time series of more than an century—we can still have an idea of the evolution over time of the proportion of doctoral students’ papers among all Quebec university papers (Figures 35 and 36). For all disciplines combined, the percentage of Quebec’s university papers authored by doctoral students increased from 20% in 2000 to 33% in 2007. The main explanation for this increase is likely the increase in the number of enrolments at the doctoral degree in Quebec, which went from 8,634 in 2000 to 11,865 in 2005 (Canadian Association for Graduate Studies, 2008). Part of this increase can also be caused by the fact that the students’ database used in this thesis does not include doctoral students who graduated or left the program in 1999 or before and who may be the authors of scientific papers written at the time they were still enrolled but published a few years later because of various publication delays. In other words, the fact that we do not have the list of doctoral students prior to 2000 underestimates the percentage of papers authored by doctoral students for years.
2000-2001, roughly\(^46\). That being said, this does not entirely explain the increase observed for recent years, which is observed in each of the disciplines of the natural and medical sciences (except chemistry) and in most disciplines of the social sciences and humanities. A similar trend was also observed by Lee (2000).

\(^{46}\) Given the standard publication delay of one or two years, it is less likely that papers published from 2002 onwards were authored by doctoral students enrolled prior to 2000.
Figure 35. Percentage of Quebec university papers to which doctoral students contributed, by discipline of the journal (natural sciences and engineering), 2000-2007
Figure 36. Percentage of Quebec university papers to which doctoral students contributed, by discipline of the journal (social sciences and humanities), 2000-2007
Figure 37 presents the cumulative distribution of the number of papers per doctoral student and faculty member, including and excluding individuals who have not published a paper over the 2000-2007 period. This Figure shows, following Lotka (1926), that for both doctoral students and faculty members, a minority of individuals are responsible for the majority of the production of new knowledge. Given the very high proportion of doctoral students (compared to faculty) who have not published at least one paper over the period, the distribution for all students is the most concentrated: 10% of students account for 70% of the papers published, 20% for about 90% and, finally, about 27% of students for 100% of the papers—that is, those who have at least one paper. The distribution of professors is also quite skewed, but not as much as that of doctoral students: about 20% of all professors account for 80% of the papers published, and about 30% account for 90% of all papers.

As established by Larivière et al. (2009), the cumulative distribution of papers is always less concentrated when only researchers with at least one paper are considered\(^{47}\). Our data confirms this observation. Indeed, the cumulative distribution of doctoral students with at least one paper is much more egalitarian, with 20% of doctoral students accounting for 50% of the papers, 40% for 70% and 70% of 90%. Because of the very high number of papers (>100) authored by a few senior researchers, the cumulative distribution of professors’ papers is more concentrated than that of doctoral students, with 20% of faculty accounting for 63% of papers,

\(^{47}\) Because of the sampling method he used—authors whose surname started with A and B and whose papers were indexed in Chemical Abstracts—Lotka (1926) could not consider authors who had no paper.
40% for 85% and 60% for 94%. Globally, when all students and all faculty members are compared, it is not a surprise to obtain cumulative distributions that are more concentrated for students: a certain proportion of doctoral students have not spent enough time in the program to publish at least one paper or do not aim at a career in research and, hence, might not diffuse their research results in a format other than the thesis48.

From a more theoretical point of view, the fact that the distribution of professors is more concentrated than that of doctoral students when only individuals who published at least one paper are considered suggests that the stratification process of science (Cole and Cole, 1973; Merton 1968; 1973)—or the formation of the elite (Zuckerman, 1977)—takes some time to arise, as the most productive doctoral students have not yet distanced themselves from the ‘average’ doctoral researchers. Although it is true that the distribution of doctoral students who are active in research is already stratified, it is not—yet—as stratified as that of professors.

48 This might especially be the case in professional disciplines such as administration and management, law, etc.
Unsurprisingly, doctoral students’ average productivity is less important than that of established researchers. As presented in Figure 38, for all disciplines combined, faculty members are more than seven times more productive than doctoral students. The difference between the two groups is higher in all disciplines of the social sciences and humanities but education, and is lower in all disciplines of the natural sciences and medicine but engineering. This is, again, likely a reflection of the different authorship attribution and collaborative practices of the different disciplines.
Data presented in the preceding section revealed that women doctoral students were less likely to publish papers than their male colleagues (Figure 17). Figure 39 confirms that this is also true for average productivity measures: in most disciplines as well as for all disciplines combined, male PhD students publish, on average, a higher number of papers during their studies than their female colleagues. Again, these measures confirms Nettles and Millets’ (2006) results obtained for a sample of doctoral researchers, as well as the results of Fox (2005); Prpić, (2002); Xie and Shauman, (1998 and 2003) obtained for more established researchers.

Several elements can explain these differences. Although we will not enquire into all of these elements—as the study of these gender differences could be a thesis in itself—let us mention a few here. First and foremost, the disciplines in which women are more likely to be active are often disciplines in which the average productivity of
researchers is less high. In the specific case of Quebec doctoral students, the fact that there is a majority of women in disciplines of the social sciences and humanities (Figure 17)—where, on average, researchers publish a smaller number of papers—might explain the difference observed at the level of all disciplines combined. That being said, it cannot explain why, for six disciplines out of nine, women are still less productive than men. Other explanations have been suggested to account for the generally observed differences in productivity, such as the lower propensity of women to collaborate (Larivière, 2007) and their lower degree of specialization (Leahey, 2007 and 2006). Although collaboration is less likely to be an issue for graduate students—as only a very small proportion of their papers are written in solo—specialization might have an effect, as doctoral students are in the part of their careers where they specialize in one area.

Several other authors have also analysed the effect of wedding and pregnancy and have found diverging results: Long and Fox (1995) and Xie and Shauman (1998) concluded that marriage can have a positive effect on productivity, although others found that children—especially when they are of young age—had a negative effect (Kyvik, 1990; Kyvik and Teigen, 1996; Stack, 2004). Long (1990) also found the same effect for graduate students. On the other hand, the presence of young children did not have the same effect on productivity for male researchers (Long, 1990).

Finally, another aspect that might affect women’s productivity during their doctoral training is their longer time in graduate school. Indeed, Xie and Shauman (1998) established that women where between 40% and 80% more likely than men to spend
more than 11 years between their undergraduate degree and their PhD and that these delays could explain their lower productivity. Although these hypotheses are all likely to influence the trends observed at the level of Quebec’s doctoral students, it is quite difficult to assess their individual effect. More micro-level data—either obtained through qualitative methods or surveys—could provide more insight into the underlying mechanisms behind the trends observed here.

5.3. Scientific impact of publications to which doctoral students contributed

Several studies have analyzed the effect of aging on researchers’ scientific impact or creativity (see the recent review by Feist, 2006). Although these studies offer diverging results, two general trends emerge from the data: older researchers publish more papers (as they are on top of the hierarchy) but younger ones typically have a higher scientific impact (as they are, it is assumed, at the top of their creativity). None
of the existing studies, however, offer any data on the scientific impact of doctoral students who are, in addition to generally being of a young age, can also be considered at the start of their academic careers. One might thus wonder, following the finding by Gingras, Larivière, Macaluso and Robitaille (2008) that younger professors’ average scientific impact is higher than that of older professors, if doctoral students’ papers are having higher scientific impact than other Quebec papers?

Figure 40 presents A) ARC and B) ARIF measures of papers to which doctoral students contributed, as well as for all other Quebec papers.\(^{49}\) Surprisingly, the indexes do not converge and reveal distinct patterns: although PhD students’ papers are generally published in journals having a higher impact factor than other Quebec papers in all disciplines combined (p< 0.001) as well as in most of the individual disciplines, their observed scientific impact (ARC) is, on the other hand, below average (p< 0.001). Students thus generally manage to publish their papers in high impact journals, but do not receive the actual citations. This suggests either that 1) there may be a Matthew Effect (Merton, 1968) for PhD students’ papers, similar to that observed for several developing countries (Bonitz, Bruckner and Scharnhorst, 1997; Pislyakov and Dyachenko, 2009) or that 2) doctoral students’ papers may be of lower scientific quality. The Matthew Effect is a sociological phenomenon observed by Robert K. Merton (1968) by which the recognition for a discovery is more easily attributed to scientists already recognized in their discipline than to others who are less known. For students—who are less known in their field than more established

---

\(^{49}\) Results labeled with *** mean that the differences are statistically significant at p< 0.001, ** =< 0.01 and * =< 0.05. The same notification is used for Figure 40.
colleagues—this means that their discoveries may be more easily overlooked than those made by better-known researchers. The observed difference between citations received and the impact factor of the journals also highlight the well-known limitations of the impact factor for predicting actual citations received (Seglen, 1992).

That being said, doctoral students’ papers nonetheless have an impact that is significantly higher than that of other Quebec papers in biomedical research, chemistry and in physics, which are also the disciplines in which their scientific papers count for the largest proportion of all Quebec papers. This suggests that increasing participation in papers on the part of doctoral students leads to better collective results: the fact that a higher proportion of papers are written by students is likely to reduce students’ Matthew Effect, as other doctoral students from Quebec or abroad are less likely to disregard papers from other students than more senior researchers in other disciplines and, in turn, more of these students’ papers will be cited. Moreover, doctoral students are expected to complete an extended and systematic review of the literature during the course of their thesis; these citations from other students might, in turn, increase the overall impact of doctoral students’ papers50. This also shows that in lab-based disciplines, technical contributions often performed by doctoral students are well rewarded (Walford, 1983), although it might not be the case in other disciplines, especially those of the social sciences and humanities.

50 Our data shows that doctoral students’ papers have longer reference lists and cite more recent papers—as measured by average age of cited literature as well as the Price Index—than other Quebec papers.
Figure 40. A) Average of relative citations and B) Average of relative impact factor of papers to which doctoral students contributed and of other Quebec papers, by discipline of the journal, 2000-2007.
As exposed in section 5.2, the vast majority of doctoral students’ papers are co-authored, although they remain the first author of the papers in more than 56% of the cases (Figure 26). Figure 41 presents A) observed (ARC) and B) expected (ARIF) scientific impact measures of doctoral students’ papers, for papers on which they are the first authors as well as for papers they are not first authors. The results are unequivocal: papers first-authored by doctoral students are less cited and are published in journals with lower impact factors than papers for which they are not first authors. As the first author is in most disciplines considered the principal author of the paper, these results may demonstrate that doctoral students’ contributions are not as significant as those to which they contributed as a secondary author. This also suggests the possible presence of a Matthew Effect, as students are usually less known than their professors. Another explanation for lower impact publications may be that doctoral students, as the first author, may opt for the journals with a lower rejection rate and a greater possibility for acceptance of their manuscripts.

As mentioned in the introduction of this thesis, doctoral students have a double responsibility. They are still in the process of learning the knowledge and *habitus* of their disciplines, but it is also expected that they contribute, at least by the time they graduate, to the advancement of knowledge in their field. This apprenticeship predictably follows an upslope curve during their doctorate degree, which suggests that researchers cannot be at their ‘prime’ years during their PhD, and hence, do not publish papers with an impact as high as that of young, yet established, researchers (Gingras *et al.*, 2008). This also suggests that doctoral students who do publish made a good choice in selecting their research teams—or that heads of teams select good
doctoral students—as the papers to which they contribute as secondary authors are more highly cited. Another possible explanation for these differences might lie in the power relationship between doctoral students and senior researchers, who might grant first authorship to students in ‘average’ scientific papers, but who might not in the case of more important discoveries.

Finally, these data show that, in today’s research system, where collaboration is the norm rather than the exception, it becomes quite difficult to attribute the paternity of an article or a discovery to an individual. Doctoral students are working in a complex system where several individuals belonging to several social groups—having their own rank in the stratification structure of the scientific community, to which a power relationship is associated—work together on a common goal (the scientific paper), but where their individual role is only imperfectly codified in the byline of papers. Although the author order provides an indication of this contribution, it does so in a limited manner, as it is a function of the power relationship between the doctoral student and more senior researchers involved in the project, but it is also dependent on the disciplinary practices of authorship attribution and name ordering (Pontille, 2004). On the whole, Figure (41) suggests that the increase in citations is an effect of the team rather than the effect of doctoral students per se, as it is known that papers authored in collaboration are more likely to be cited (Glänzel, 2001; Iribarren-Maestro et al., 2009; Lawani, 1986).
Figure 41. A) Average of relative citations and B) Average of relative impact factor of papers to which doctoral students contributed, either as first author or other author, by discipline of the journal, 2000-2007
Research in the psychology of creativity often shows that individuals are in their ‘primes’ when they are younger (Simonton 1984 and 1997). One might then argue that, although doctoral researchers’ papers are less cited than other Quebec papers, they are more likely to represent breakthrough discoveries. Figure 42 provides evidence for the opposite: for all disciplines combined, papers to which doctoral students contributed are, on average, less likely to be among the top 5% most cited papers of their respective disciplines. We nonetheless measure variations among disciplines: for five disciplines of the natural and medical sciences (biomedical research, chemistry, engineering and technology, mathematics and physics) and for two disciplines of the social sciences and humanities (health and social sciences), doctoral students’ papers are more likely to be among highly cited papers.

Figure 42. Percentage of doctoral students’ papers and of Quebec other papers among the top 5% most cited papers, by discipline of the journal, 2000-2007
Figure 43 provides a gender breakdown of doctoral students’ scientific impact. Although the Figure shows differences between male and female students papers’ scientific impact, Mann-Whitney statistical tests revealed that none of these differences were statistically significant. Although previous studies have provided evidence of differences in citation rates (Turner and Mairesse, 2005; Peñas and Willett, 2006), several other studies have obtained similar citation rates for both men and women (Bordons et al., 2003; Gonzalez-Brambila and Veloso 2007; Long and Fox 1995; Mauleón and Bordons, 2006). Our results provide additional support for the latter studies; the gender effect on scientific impact found in former studies could be caused by the fact that the authors have not normalized citation rates by the specialty of the paper and that women are more likely to be active in these lower impact areas.
This chapter provided the first large-scale empirical data on doctoral students’ contribution to scholarly research. It revealed that a large proportion of the province’s papers have students as authors (~30%), but that this percentage varies a great deal among the various disciplines. More specifically, it is higher in the disciplines of natural and medical sciences (except mathematics), and lower in the disciplines of the social sciences and humanities (except psychology). Several elements can explain these differences, among which labor division and collaboration have a central importance. Our data also uncovers that the share of Quebec’s papers to which doctoral students contributed increased over time, which is likely a reflection of the increase in co-authorship observed in most of the disciplines (Larivière, Gingras, Archambault, 2006). The distribution of the number of papers by doctoral student also established that the stratification structure of doctoral students
is much more egalitarian than that of established researchers. This indicates that it
takes some time for researchers to accumulate publications and that the ‘elite’ of
young researchers cannot yet distance themselves from the average researcher during
their doctorate, at least in terms of papers published. In other words, because of the
lower number of papers published by doctoral students—even the most productive
ones—the elite of doctoral students do not contribute to a proportion of papers that
is as large as that of the elite of professors in all professors’ papers.

Data presented in this section also provided evidence of a gender gap in doctoral
students’ productivity, with males producing, on average, a third more papers than
their female colleagues. These differences in productivity remain even within the
same discipline and several factors—other than their choice of research topics—are
generally mentioned in the literature for explaining these differences, such as
marriage, pregnancy, etc. Further research is needed on this aspect to assess whether
these factors affect Quebec’s PhD students as well. On the other hand, no significant
gender difference has been observed in terms of scientific impact, a finding that is
consistent with most studies (Bordons et al., 2003; Gonzalez-Brambila and Veloso
2007; Long and Fox 1995; Mauleón and Bordons, 2006).

This chapter also investigated whether young researchers have more scientific impact
that older ones, with a mixed result. Even though doctoral students’ papers are
generally published in journals with impact factors greater than that of other Quebec
papers, the opposite is found when observed citation rates are compiled. This suggests
that doctoral students’ papers may suffer from the Matthew Effect (Merton, 1968).
The disciplinary breakdown in terms of *expected* vs. *observed* impact also provided evidence that this effect is less strong in the natural and medical sciences than in the social sciences and humanities. Indeed, in the former group of disciplines, a higher proportion of doctoral students are the authors of papers and, since we can assume that they are less likely to ignore papers authored by their student colleagues, the citation rate of doctoral students is higher.
A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents die, and a new generation grows up familiar with it [Physicist Max Planck, 1858-1947]

6. Conclusion, limitations and further research

As the first large-scale study of a complete population of doctoral students’ scientific publication, this thesis achieves three main contributions to the advancement of knowledge in the fields of information science, sociology of the scientific community and sociology of higher education.

The first contribution is a technical one and involves the creation of an algorithm that allows the automatic attribution of a large proportion of individual researchers’ papers. The attribution of individual researchers’ papers is a complex task because of the limitations of bibliographic databases as well as the existence of homographs, i.e., researchers having the same name. Using an already-cleaned WOS-indexed publications list for all university professors in the province of Quebec, we searched for patterns which could be used to automatically assign a large proportion of the papers of another group of researchers: doctoral students. Two types of patterns were found: 1) at the individual researchers’ level and 2) at the level of disciplines.

At the level of individuals, we found that researchers were quite regular in their referencing practices. This could be expected: as shown elsewhere, researchers tend to cite the same material throughout their careers (Barnett and Fink, 2008; Gingras et al., 2008). We thus tested this finding for the subset of Quebec researchers and found that papers with the same surname and initial were generally those of the
‘right’ researcher when at least one of the references of the paper had already been made in one of the papers previously assigned to the researcher. Similarly, researchers also tend to work on the same topics. Using the pool of keywords previously used by researchers and comparing them with papers subsequently published, we found that the use of the same keywords meant in most of the cases that the paper belonged to the same researcher.

At the collective level, two general patterns emerged. The first pattern we found was that the institution of affiliation of a given researcher appeared on most of the papers that rightly belonged to him/her. This simple regularity allowed the creation of a core subset of papers, which could then be used to gather the researchers’ other papers using the previous references and previous keywords methods. The other pattern relates to the relationship between the department discipline and the speciality of the journal in which papers are published. For some departments/specialty combinations, a majority of papers belonged to the ‘right’ researcher, while for other combinations, a majority belonged to homographs. Thus, the former combinations allowed the automatic attribution of papers, while the latter made automatic rejection of author-paper combinations possible.

The combination of these patterns allowed the automatic attribution or rejection of at least one paper for more than 88% of the 27,397 doctoral students. It also automatically closed 9,595 files out of the 15,702 for which at least one author matched the name of a doctoral student. Out of these 9,595 files, 5,352 files had all of their papers rejected and 4,243 had all of their papers assigned. Compared with
most existing studies on author disambiguation, which were generally performed for a small subset of researchers (Han et al., 2005; Aswani, Bontcheva and Cunningham, 2006; Wooding et al., 2006), for specific author-article combinations (Boyack and Klavans, 2008), and often without a test bed for assessing the error rate of their algorithm, this is an important step forward. That being said, the recent developments in bibliographic databases used in bibliometrics—such as the link between each of the authors and their addresses as well as the indexing of the complete given names of authors—are perhaps even more important, as they are likely to make this attribution work easier.

The second contribution of this thesis is its unique large-scale empirical data on the socialization of doctoral students to research, as evident through publication records, with respect to a complete population. Previous studies mainly explored this process from a qualitative standpoint which, by definition, uses a small sample of students in one or a few disciplines or universities. This thesis is at the other end of the spectrum: it analyzed this process from a quantitative—even epidemiological—point of view, using a simple, yet relevant, indicator of socialization to research: the participation in peer-reviewed papers. Using this indicator, this thesis investigates, for the complete spectrum of academic disciplines, 1) the extent to which doctoral students are socialized and integrated into research, 2) the context of this socialization and 3) the relationship between this socialization and their subsequent research careers.
The results show that the percentage of doctoral students who are involved in research—as measured by their involvement in publications—varies a great deal among disciplines, with a high of 80% in basic medical sciences, but a low of 10% in education. Although the disciplinary classification used in this thesis is not exactly the same as that of Nettles and Millets (2006) we see that the ‘spectrum’ of doctoral students’ participation in papers by discipline is quite similar, with students in the sciences publishing more papers and students in education publishing less papers. A similar trend is observed in Lee’s (2000) data.

Several factors may explain these differences, among which the different formats of doctoral theses (article-based vs. monograph) as well as the various modes of organization of research are probably the most important. Although we do not have data on the extent of article-based theses and of its differences among disciplines, we know that this ‘form’ of thesis has become quite frequent in the natural and medical sciences (Breimer, 2010; Holdaway, 1994). And this may affect the contribution of doctoral students to papers in these disciplines.

Along the same lines, in the natural and medical disciplines collaboration is the preferred mode of production of knowledge (Larivière, Gingras and Archambault, 2006) and doctoral students’ research projects are de facto linked with those of their supervisors (Gemme and Gingras, 2008). Doctoral students, then, become part of their advisors’ research team and work, along with the other members of the team, on the various projects of the lab, of which their own doctoral research is only one component (Pole et al., 1997). This integration in a research team takes away the
burden of socialization and integration on the sole shoulders of the advisors, as several other actors (postdoctoral fellows, laboratory technicians, etc.) are involved in the process (Delamont, Atkinson and Parry, 1997; Gemme and Gingras, 2008), and increases the opportunity for doctoral students to participate in various research projects.

However, in the social sciences and humanities—where research teams are not the norm (Larivière, Gingras and Archambault, 2006)—doctoral students are less integrated into their advisors’ research (Delamont, Atkinson and Parry, 1997; Ridding, 1996), as professors are either less likely to need help in their research or the—often technical—work the students perform does not grant authorship. They typically work as assistants on their advisors’ projects—which may or may not be linked with their thesis—for a few days per week and then work in their own projects the rest of their time (Legault, 1993).

One might argue, thus, that students in the social sciences and humanities are as integrated into research as their colleagues in the natural sciences and that the bibliometric differences observed in this thesis only reflects the different authorship practices. This seems like a remote possibility, however, as these authorship practices are also a reflection of the different type of technical work that is performed in these disciplines. Indeed, while technical work in the natural and medical sciences involves the manipulation of technological artefacts that requires important scientific knowledge on the part of the experimenter—such as microarray devices, telescopes, spectrometers, etc.—, the tools used in social sciences and the humanities are
generally less complex and do not need students to possess an advanced technical knowledge. Further research is needed in this area to assess the ‘complexity’ of the task performed by graduate students, and its effect on their role as co-authors.

The access to research funds also influences doctoral students participation in peer-reviewed papers, as specialties where professors receive major research funding are also those in which students are the most likely to publish. This finding is consistent with those of Nettles and Millets (2006) and Buchmueller, Dominitz and Hansen (1999). Major research grants received by faculty members often serve to grant scholarships or to hire students to work on their research projects. Students might, thus, have more opportunities to participate in the publication of peer-reviewed papers.

Following Nettles and Millets (2006) and Paglis, Green and Bauer (2006), collaboration could be considered as the common denominator of the socialization process: in all disciplines of both the natural and medical sciences as well as in the social sciences and humanities, doctoral students’ papers are more often written in collaboration than are other Quebec papers. On the other hand, a smaller proportion of doctoral students’ papers, in comparison with other Quebec papers, are authored with foreign partners, which may suggest that doctoral students’ socialization happens in a more local setting. Finally, being socialized to the publication habitus during the doctorate is associated with graduation and post-graduation research activities, as those who are more active in research publications during their PhD are
more likely 1) to graduate and 2) to continue being active in research after graduation.

The third contribution of this doctoral research is to measure the contribution of the research output produced by doctoral students for the research system. We knew that PhD students account for a high proportion of the academic workforce in terms of numbers of individuals (Canadian Association for Graduate Studies, 2006). This thesis shows that they also contribute a relatively large part of the overall scientific output, 33% in 2007. This percentage is slightly higher than that obtained by Cursiefen and Aktunbas for medical students in Germany (28%) and by Whitley, Oddi and Terrell (1998) for students in nursing (31.6%).

But how significant is this output? The unique aspect of this dataset makes comparisons quite difficult. Doctoral students do not publish as much as faculty members, a finding which may be explained by access to resources and their shorter research experience. No other large scale study has been done on the topic in other provinces or countries and, hence, no comparable data exists in the literature.

One can, nonetheless, get an idea of the extent of doctoral students’ scientific output by weighting the percentage of Quebec papers to which doctoral students contributed against the percentage of other research sectors’ publications in the overall Quebec output. Although students cannot be considered as a ‘sector’ from an organisational point of view, this comparison can nonetheless provide a ‘rough’ sense of the size of doctoral students’ contribution to the output of the research system.
This comparison emphasizes the fact that doctoral students contribute to more papers than do all of Quebec hospital researchers combined, and five times more than do federal or industrial researchers. The contributions of these sectors in the province’s output being generally acknowledged as important (Godin and Gingras, 2000a), we can argue that doctoral students’ contributions in the system is at least as central as that of hospitals, and of governmental and industrial laboratories. This obviously does not suggest that PhD students are as important to the research system as hospitals or industries—as these sectors perform specific tasks which could not be performed by students alone. But still, the comparison shows that the output of students, as measured by the publications to which they contribute, is larger than the publication output to which hospitals or industries contribute.

In terms of scientific impact, things are much more complex. Although doctoral students’ papers are published in journals with significantly higher impact factors than other Quebec papers, they obtain citation rates that are significantly lower. This finding may be due to the Matthew Effect (Merton, 1968). Students are less known than more established researchers and, hence, their papers may be more easily overlooked by other researchers. Another factor which might explain PhD students’ lower citation rates is their lower international collaboration rates, as papers written in international collaboration obtain, on average, higher citation rates (Glänzel, 2001). These scientific impact measures also complement findings in recent literature regarding the effect of aging on the scientific impact of Quebec professors, which showed that, for all disciplines combined, the average scientific impact of researchers’ papers decreases with age (Gingras et al., 2008). Our data on doctoral
students suggests either that 1) before decreasing, the average scientific impact of researchers’ papers increases—as they gain more research experience—until they are hired as professors or 2) that the subset of doctoral students with the highest scientific impact have more chances of becoming faculty members. Scientific impact data also provide evidence of a strong disciplinary component to this scientific impact or ‘creativity’, as students’ papers in empirical disciplines obtain higher citation rates while those in more theoretical disciplines obtain lower citation rates.

Our analysis also establishes gender differences in favor of males both in terms of the percentage of students having published papers as well as the average number of papers published. This is consistent with the findings of most authors (among others Fox, 2005; Prpić 2002; Xie and Shauman 1998 and 2003; Zuckerman, 1991); our analysis provides evidence that these differences are already present at the doctorate level. On the other hand, our data reveals no significant difference in terms of specialty-normalized scientific impact, which suggests that the differences often observed in that respect are caused by women’s research domains; meaning that the domains in which women publish are generally less citation intensive than those in which men are most active.

Finally, our analysis also highlights, in today’s highly collaborative research environment, the difficulties of attributing the credit of a scientific discovery to a single individual. This is a problem not only for our study, but for bibliometric research in general. Although the first author of a paper is generally considered the main individual responsible for the research, it might not always be the case because
of 1) the power relationship between the doctoral student and more senior researchers involved in the project as well as 2) the disciplinary practices of authorship attribution and ordering. As only a handful of studies have looked at authorship attribution methods in science (Biagioli and Galison, 2003; Birnholtz, 2006; Pontille, 2004), it would be very interesting to see in further research if there are differences in authorship attribution methods in the case of faculty-student collaboration. Although these qualitative studies on authorship provide evidence of the existing practices in several disciplines, we do not have the full spectrum yet. There might also be a local aspect: other countries, for example, might have different authorship attribution practices\textsuperscript{51}. More generally, this also underlines the need for studies on what a contribution to a paper is and on the different types of contributions that authorship attribution practices imply.

This research also has other limitations. Because of the single indicator used—peer-reviewed papers published by doctoral students indexed in the WOS—, it only provides a partial view of doctoral students’ socialization to research, as this socialization process also takes other forms, both in terms of research outputs (books and book chapters, conference proceedings, etc.) and also in terms of other activities which might not necessarily lead to a publication output (participation in seminars and in international conferences, etc.). In other worlds, only contributions

\textsuperscript{51} A quick analysis of the World’s most productive researchers showed that they were mostly from Asian countries—even after the manual cleaning of homographs—which is likely a function of their hierarchical structure and its effect on the “automatic” attribution of authorship to senior researchers in charge of the lab, irrespective of their actual contribution to papers.
that lead to the writing of WOS-indexed peer-reviewed papers are measured in this thesis.

Similarly, this study does not provide any insight into the status of the other authors of papers, such as technical staff, masters’ students and postdoctoral fellows. It could indeed be expected that postdoctoral fellows account for a very important proportion of the province’s papers, and that this social group, being closer to their doctoral years than established researchers, have an important role in doctoral students’ socialization to research, a finding which has been previously reported by Gemme and Gingras (2008) for a sample of doctoral students in Quebec. These are empirical questions that can be answered only with empirical data; and it is hoped that further studies will provide such data. Similarly, it would also be relevant, for further research, to explore in depth doctoral students’ trajectories after they graduate and to study the characteristics of the scientific production of those who become research professors. For example, we could explore whether students with high scientific production and scientific impact have a greater chance of becoming faculty members. The dataset used in this thesis could also be complemented with data on funding received by students via federal and provincial research councils—the only consistent source of data on graduate students’ scholarships—to measure the effect of direct funding received by students on the trends observed. Along these lines, further quantitative research in other countries could shed light on the universal aspects of doctoral students’ participation in research, as well as on those that are culturally specific.
Given the fact that, by definition, qualitative studies on this process provided more narrow—albeit often richer—results, the combination, in further research, of both quantitative and qualitative approaches to the study of socialization is likely to provide interesting results. Further research in the area could also combine this quantitative data with a qualitative analysis of the factors that affected PhD students’ contribution to papers. For instance, it would be interesting to conduct interviews with these four possible cases: 1) students who graduated but did not publish papers; 2) students who graduated and published; 3) students who did not graduate but published; 4) students who did not publish nor graduate. These interviews could also shed light on the ‘nature’ of the student contribution that lead to authorship credit on a paper and, on the other hand, on other types of contributions to research which did not lead to authorship of papers. More generally, research on the nature of the work conducted by doctoral students working as research assistants would also contribute to the understanding of the mechanisms behind the trends observed in this thesis. Similarly, interviews with supervisors—who generally are the ‘grantors’ of authorship—on the factors affecting students’ participation in the publication process and on their ‘norms’ for granting authorship are deeply needed.

Overall, this interdisciplinary thesis—which combined advanced bibliometric measures with Bourdieu (2001) and Merton’s (1973) sociology of science—makes an original contribution to information science, taking into account the central importance of bibliometric methods, and to the sociology of science and of higher education. The results obtained provide important insight into the extent, the context and the effect of socialization to research in the PhD curriculum, as well as a better
understanding of the importance of doctoral students’ scientific contributions within Quebec’s research system, both in terms of papers and citations. More generally, it also shows the importance of doctoral programs, not only for the ‘reproduction’ of new researchers, but also for the research system as a whole, as PhD students contribute to an important proportion of the knowledge output of system during their apprenticeship. These findings should be of great interest to university administrators as well as to research councils and the science policy community in general.
Bibliography


208


