Researchgate.net Crawler and a new Contribution Determines Sequence (CDS) Method

by

Zahra Hammook

BSc, Garyounis University, 2002

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Science

in the Program of

Computer Science

Toronto, Ontario, Canada, 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Researchgate.net Crawler and A new Contribution Determines Sequence (CDS) Method

Master of Science 2015

Zahra Hammook

Computer Science

Ryerson University

Abstract

General and Focus crawlers are the main types of web crawlers used for different goals, with different crawling techniques and architecture. Our crawler was written in Java language using different software and libraries. To test the crawler, it has been run on the academic social network, Researchgate.net from 3^{rd}.April to 28^{th}.June 2014 and retrieved real data. The crawler consists of three main algorithms to crawl information such as researchers details, publications details, questions/answers activity details. The retrieved data has been analyzed to highlight the performance of Canadian researchers, in the field of Computer Science on Researchgate.net. Data analysis has been done from the collaboration and (alt)metrics perspectives. Among other features Researchgate.net came with "Impact Points" and "RG Score" (alt)metrics. The former builds on ISI Journal Impact Factor, which disregards author's contribution in its calculations. A new Contribution Determines Sequence (CDS) method has been developed and tested, with all required scripts which showed better performance than other methods.

Acknowledgements

I would like to express my special appreciation and thanks to my advisor Professor Dr. Jelena Misic, who has been a tremendous mentor for me. I would like to thank her for encouraging my research and without her care and patience I would never have been able to finish my thesis. Her advice on both research as well as on my career have been priceless. I am fortunate to work under her supervision, which contributed much to my knowledge and research experience. I am fortunate to be supervised by such a knowledgeable professor like her. Many thanks and deep gratitude to you.

I would like also to thank my committee members, for their brilliant comments and suggestions. Thank you for serving as my committee members and making my defense an enjoyable moment.

A special thanks to my family. Words cannot express how grateful I am to my mother, and father, for all of the sacrifices that they made on my behalf. Their prayers for me were what sustained me thus far and incited me to strive towards my goal. At the end I would like to express appreciation to my beloved husband who spent significant time caring for our family during his vacation with us. Deeply grateful for all of your support.

Contents

Declaration ii

Abstract iii

Acknowledgements iv

List of Tables ix

List of Figures xi

1 Introduction **1**

1.1 Crawler Identification 2

1.2 Crawlers and Crawling Social Networks 3

1.3 Academic Social Networks 5

1.3.1 Social Networks 5

1.3.2 Academic Social Networks 5

1.4 Research Methodology 7

1.5 Research Questions 8

1.6 Research Contributions 8

1.7 Research Organization 9

1.8 Summary 9

2 Related Work **11**

2.1 Summary 16

3 Crawling Researchgate **17**

3.1 Crawler (Overview) 17

3.1.1 Queues : 17

3.1.2 Crawling Loop : 18

3.2 Crawler's Features 19

3.3 Software Needed 19

3.4 A crawler Architecture 20

3.5 Main Algorithms 21

3.5.1 Crawling Researchers' Information 22

3.5.2 Crawling Publications Information 28

3.5.3 Crawling Information on Questions & Answers 32

3.6 Classes Used 34

3.7 Summary 36

4 Testing the Crawler **37**

4.1 Collaboration-Co-authoring 38

4.1.1 Co Authoring between Supervisors and Students (MS.c.) 38

4.1.2 Co-authoring between Supervisors and Students (Ph.D.) 41

4.2 Co-authoring among Canadian Researchers: Locally, Nationally and Internationally 46

4.3 Knowledge Sharing among Canadian Researchers on Researchgate.net 49

4.3.1 Who Answers on Researchgate.net? 50

4.4 Researchgate.net Metrics 56
4.4.1 Correlation between Followers and Downloads 56
4.4.2 Correlation between Followers and Views 61
4.4.3 Correlation between Publications and Views 62
4.4.4 Correlation between "Views" and "Number of Authors Per Paper" 64
4.4.5 Members of Department versus Authors of Publications 65
4.4.6 Correlation between Departments Publications and the Total Number of Impact Points 66
4.5 Altmetrics Influence 68
4.6 Metrics Changes 71
4.6.1 "Followers" and "Publications" on Reserachgate.net 71
4.6.2 "Views", "Citations" and "Downloads" on Researchgate.net 72
4.6.3 "Impact Points" and "RG Score" Change on Researchgate 72
4.7 Author's Position 73
4.7.1 The Relationship between Total of Publications and 1st Co- Author . . . 75
4.7.2 The Relationship between Total of Publications and 2nd. Co- Author . . 76
4.7.3 The Relationship between Total of Publications with 3rd Co-author . . . 77
4.8 Summary 78

5 Impact Points and RG Score **79**

5.1 Author Listing on the Byline 80
5.2 Suggested Method 81
5.2.1 CDS Method Features 83
5.3 Credit Allocating Schemes 85
5.3.1 The Simplest Equalitarian Fractional Allocating 85

5.3.2 Tscharntke, Teja Scheme 85
5.3.3 Arithmetic Allocating Scheme 86
5.3.4 Geometric Allocating Scheme 86
5.3.5 Tailor Based Allocations(TBA) 87
5.3.6 Suggested Contribution Determines Sequence (CDS) Method 89
5.4 CDS Calculation Scripts 91
5.5 Experiment 94
5.6 CDS Main Advantages 97
5.7 Summary 97

6 Conclusion and Future Works 99

Appendix 102
A Crawler's Algorithms

References 119

List of Tables

4.1 First and second moments along with the variance and Stdev for Ph.D students/supervisors/co supervisors publications. 42

4.2 The percentage of actual and joint publications. 43

4.3 First group ranging from (1-800) of collaboration on “Local” , “National”, “International” levels. 46

4.4 Second group ranging from (801-1600) of collaboration on “Local” , “National”, “International” levels. 47

4.5 Third group ranging from (1601-2400) of collaboration on “Local” , “National”, “International” levels. 47

4.6 Fourth group ranging from (2401-above) of collaboration on “Local” , “National”, “International” levels. 47

4.7 Question/Answer on Researchgate.net. 49

4.8 Answers from Canadian researchers and different countries 51

4.9 Performance of different formats materials upload on Researchgate.net. 55

4.10 Finding r between “Followers” and “Downloads”. 58

4.11 Analysis of Variance “Followers” and “Downloads”. 58

4.12 Finding r between “Followers” and “Views”. 61

4.13 Analysis of Variance “Followers” and “Views”. 62
4.14 Correlation between “Publications” and “Views”. 63
4.15 Analysis of Variance “Publications” and “Views”. 63
4.16 “Followers” and “Publications” readings. 71
4.17 “Views”, “Downloads” and “Citations” fluctuation 72
4.18 “Impact Points” and “RG Score” on Researchgate.net. 73

5.1 Different allocating schemes results proposed by Tscharntke. 84
5.2 Suggested approach (method) for three different reading 8, 12, 16 Co Authors. 84
5.3 Dividing impact factor by using Fractional Allocating Scheme. 85
5.4 Dividing impact factor by using Teja Tscharntke Scheme. 85
5.5 Dividing impact factor by using Arithmetic Allocating Scheme. 86
5.6 Impact factor reading by using Geometric Allocating scheme. 87
5.7 Impact Factor by using TBA scheme. 88
5.8 Results of Suggested Contribution Determines Sequence Scheme 93
5.9 Suggested Contribution Determines Sequence Method is applied to five different researchers with different positions. (Ranking first initials were used for privacy issues) 96
5.10 Comparison among schemes reading and finding the closest one to Researchgate.net reading. 96

List of Figures

3.1 Crawling Process. 18

3.2 Researchgate.net Crawler Architecture. 21

3.3 A clarification of how to match the data. 23

3.4 A clarification of Fetching Departments of Computer Science. 24

3.5 Researchers Crawler Process and Data Obtained as .csv Format. 25

3.6 A heritage Algorithms Tree. 26

3.7 Crawling Publication Details and Data Obtained as a .csv Format. 31

3.8 Crawling Questions/Answers Details. 35

3.9 Questions/Answers Data Obtained as .csv Format. 36

4.1 Supervisor/Student (MS.c.) co-authoring. 39

4.2 Euler and Venn Diagram. 40

4.3 Euler and Venn diagrams for joint supervisor/student(MS.c.) publications. . . . 41

4.4 Euler and Venn Diagram for joint publications among Ph.D students, principal supervisors and co-supervisors. 45

4.5 Countries ranking in terms of Question/Answer activity. 54

4.6 “Followers” versus “Downloads”. 61

4.7 “Followers” versus “Views”. 62

4.8 "Views" versus "Publications". 64
4.9 "Views" versus "Authors Per Paper". 65
4.10 "The Number of Department's Members" versus "The Number of Authors of Publications of Computer Science Department". 66
4.11 "Publications" versus "Total Impact Points". 67
4.12 Developments of Departments altmetrics. 68
4.13 Box Plot for "Downloads" and "Researchers' Ranking". 69
4.14 Data Categorized based on The Number of Publications. 70
4.15 Data Categorized based on The Number of Citations. 70
4.16 Data Categorized based on The Number of Followers. 71
4.17 Correlation between Publications and Main author position. 75
4.18 Correlation between Publications and 1st. Co-author. 76
4.19 Correlation between total of Publications and 2nd Co-author. 77
4.20 Correlation between total of Publications and 3rd.Co-author. 78

5.1 Structure of Contribution Determines Sequence (CDS) Scheme. 90
5.2 Pseudo Code for Calculating The IF for Group1. 91
5.3 Pseudo Code for Calculating The IF for Group2. 91
5.4 Pseudo Code for Calculating The IF for Group3. 92
5.5 Pseudo Code for Calculating The IF for Group4. 92
5.6 Pseudo Code for Calculating The IF for Group5. 92
5.7 Suggested Contribution Determines Sequence Compared to Other Schemes. . . . 93
5.8 Illustrative Basic Research Components on Contribution Determines Sequence (CDS) Method. 95

Chapter 1

Introduction

Crawlers and crawling are two terminologies that are widely used in the field of computer science. Basically, crawling is a procedure through which web pages are gathered. Web pages are gathered with the help of hyperlinks that are followed. Normally, this is done with a small set initially and then a larger set is followed in order to gather large number of web pages. After web pages are gathered, further processing is done. A suitable example to quote here is of a web search engine that has gathered several pages before the preparation of an index, in order to make them available for customers or people. Similarly, another example is of a crawler that utilizes or follows a set of such web sites that are social networking websites. Since crawling is the terminology that gathers hyperlinks, the whole procedure is completed with the help of crawler. Crawler is a program that is normally known as a robot, spider, or even an a bot. Crawler is a simple program that uses a seed list. A seed list is the list of pages from where the process or collection of web pages start. Crawler developed with the development of the Internet and the challenge of huge data on the Internet and interactivity that led to the development of Deep Internet and Rich Internet Applications (RIA) crawlers [1]. Presently there are many types of crawlers to meet the constant change on the Internet. Incre-

mental crawler is a traditional one which to refresh its collection, replaces the old documents with newly downloaded ones. The advantage of incremental crawler is that the user is provided with valuable data, achieving data enrichment and maintaining network bandwidth. A Focused Crawler or topical crawler, downloads related pages determining way forward relevancy. It is economical on hardware and the network resources. A Distributed crawler applies distributed computing techniques for extensive web coverage using Page rank algorithm. The main advantage of this crawler is flexibility. A Parallel Crawler depends on Page freshness and Page Selection allowing for multiple crawling by running many crawlers in parallel (C-Procs) [2]. Development of a suitable or effective crawler is not an easy job as there are number of challenges that subtly interfere and create issues, especially in large scale web crawlers. As mentioned, there are numerous challenges, however, some to mention in this context are, politeness for the web servers, duplicate detection, URL normalization, queue maintenance of un-fetched web pages, re- crawling as well as to prevent spider traps. On the other hand, in case of large scale crawlers, throughput increment and resource utilization are the main issues that have to be managed in order to liberate coverage [3].

1.1 Crawler Identification

In order to gather crawlers, it is important that the web crawlers identify each other. In this way, similar kinds of web pages will be gathered in a location. Normally identification of crawlers is done with the help of HTTP requests user- agent field. With the help of user agent field in the HTTP request, the web administrators are able to identify which and what type of web servers have been visited as well as their frequency. Along with this, the user agent field of the HTTP request is usually capable of providing the crawlers information. Therefore, it is a benefit for the web site administrator to look out for the URL which may provide some extra

information about the crawler [4]. Another important aspect about crawlers is that they should identify themselves. This is so because, web administrators can contact the web site owner in case of any issue/s. Social networking crawlers gather as many pages as possible that take a person directly to the social networking websites. There are numerous kinds of social networks such as Facebook [5], LinkedIn [6], Twitter [7], and many others.

1.2 Crawlers and Crawling Social Networks

Online social networks have become very much popular and due to their popularity, there is a huge increase in various data collection platforms. Social networks have become a platform for people to share information and communicate with people far away [8]. In correspondence to a crawler in social network, it can be termed as a program that initiates with the most visited page by users. With the help of crawlers, crawling is performed to retrieve information from the social networks. There are different kinds of crawlers that are used for this purpose. Parallel crawlers is a simple example of this fact. With the increase in web sites size, the need of web crawlers also increase, so that huge amounts of data can be stored. In a study, eBay was used as an online social network from where the user profiles were retrieved. With the help of crawlers, personal information about the users can be retrieved such as their name, contact information or pictures and videos. Since, crawling is a program that is used for the collection of web pages, there are different factors that are being evaluated while a crawler is developed. The factors included are: a-Selection of seeds: As seed is the initial point from where crawling starts, therefore, selection of seeds should be done carefully in order to avoid low quality web-sites to be included in the search engine list, b-Selection of node algorithms: Node algorithms are those algorithms that decide which website should be displayed in the search engine list next. The most common example of node algorithms is breadth- first search (BFS) c-Users

protection: Since huge amount of data is stored when large number of users are present, it may be possible that the crawler misses out some significant amount of searches. Therefore, it is important to consider protection of users profile information or personal information Properties of Online social networks: This is another important aspect as online social networks are numerous and all have different features and properties. Therefore, prior to the development of crawler, properties of online social networks have to be checked thoroughly

d-Sensitivity: This factor evaluates how crawling is affected due to online social networks and black hole users e-Bias: In this, the crawling sub graphs developed from the whole graphs are evaluated with the help of statistical properties f-Efficiency: This is also important as it describes how well and how fast the search engine might reply or respond [9]. We tested our crawler by crawling the academic social network Researchgate.net. Early academic social networks Mendeley, Zetro and Connotea, were mainly reference manager tools that enabled researchers to share their references [10]. A researcher can import/export citations, and generate bibliographies automatically [11], allowing him to list his publications on his profile. Today academic social networks are on the rise and Researchgate.net and Academia.edu are good examples. While the focus of reference sharing sites is on enabling readers to find and share references, Academia.edu and Researchgate.net focus is on researchers themselves and their contribution enabling academics to create their own profiles with personal information, research interests, allowing a researcher to follow or be-followed, making /answering questions and reporting on user activities. These new platforms enabled researchers to communicate, collaborate and follow each other, permitting easier knowledge sharing [12].

1.3 Academic Social Networks

1.3.1 Social Networks

Social Networks like Facebook and Twitter are well established and much used for exchanging general ideas, photos and communication, but its features make it attractive to be used by researchers as well. Twitter [7] with its hashtag feature [13] is used by scientists and academics for social interaction and scientific purposes such as scholarly conferences, where academics are able to communicate easily about conferences or other topics [14]. Facebook [5] has been used by academics for various scientific communication and information sharing in addition to its main purpose as a social tool. While Facebook is mainly a social network, it has good potential of being used for focused groups of the academic community for academic purposes [15]. LinkedIn [6] is rather a profession- oriented social network, but it has added recently a new analytical tool to its publishing platform, allowing authors to track traffic received by their posts [16].

1.3.2 Academic Social Networks

Academic Social Networks are used by scholars for communication and research-related purposes. Different from the earlier academic social networks like Mendeley [11], Zotero [17] and CiteULike [18] which were meant for uses as references and files sharing. The recent academic social networks, Academia.edu and Researchgate.net came as full collaboration platforms. They allow users to communicate, collaborate, and follow or being followed, attracting millions of researchers providing better channels for scholarly communication. The declared target behind launching Academia.edu by an Oxford University philosopher 2008 was to serve as academic social network connecting researchers and allowing for information sharing and exchange. Researchgate.net, incepted in the same year (2008) by a physician, witnessed a viral expansion

with more than 5 million members presently [19]. It has features similar to Academia.edu with other features borrowed from Twitter and Facebook, emphasizing discussions and collaboration. Researchers on Researchgate.net can create and modify their profiles, upload/download publications, view, comment, make/answer questions, follow or being followed by RSS service to keep current and up-to-date. The main advantages of Researchgate.net to researchers, is that it allows self-archiving, reputation building and informal exchange of publications, which would result in better publication visibility and knowledge sharing. It is possible to find papers from within Researchgate.net and, to search some external databases such as PubMed, CiteSeer and arXiv through its efficient search engine. It is easy for a researcher to advertise on his profile different events such as meetings and workshops. The network provides a platform for a researcher to create a profile, publish his/her papers and communicate with other researchers, presenting a new way for scholarly communication. Resesearchgate.net is receiving good attention and becoming popular among the researchers community. Its popularity ranking is shown by Alexa.com, which calculates the global rank of a website using a combination of average daily visitors and page views on the site for the last 3 months. The rank jumped from 3,947 by Nov.2013 to 1,385 by 22nd.Feb 2015 to 1,194 by 21th.Mar. 2015 [20]. Researchgate.net intake is increasing and the academic community seems to be interested in this academic social networks, using it for different purposes, with some noticed "User resistance" to using academic social networks generally [21]. Actually Researchgate.net and other academic social networks created a new way of publications dissemination and communication between researchers, which would, hopefully, result in more collaboration, knowledge sharing and open access to scientific literature, supplementing universities efforts to increase researches visibility by creating repositories [22]. This can alleviate fee-based access to scientific publications [23]. Since academic social networks offer an attractive alternative to meet researchers' information needs in addition

to other features where even negative results can be reported. When Researchgate.net reaches satisfactory level of intake of academic faculty members, we think that it has the potential of becoming one of the evaluating tools of researchers performance in the future. Researchgate.net platform has not been investigated except by a bunch of limited studies and highlighting collaboration among Canadian researchers on different levels is important to understand the research dynamics among them in the field of computer science on Researchgate.net.

1.4 Research Methodology

To test the crawler it was implemented for crawling Researchgate.net. A bi-weekly run of the crawler was made for three months starting from 3^{rd}.April to 28^{th}.June 2014. Limited studies have been done on crawling Researchgate.net and since, dealing with a dynamic academic social network, a researcher's presence on Researchgate.net is not regular and statistics change. The crawler retrieved (1200) out of the total of (1563) and a sample drawn of (506) Canadian Computer Science researchers from (32) Canadian Universities.

Three readings have been collected for each researcher using colour coding in MS Excel and fifteen to sixteen records picked for each university depends on the total number of members from that university having an account on Researchgate.net. Data has been analyzed,and interpreted, incomplete profiles were discarded. Different sources have been consulted such as Canadian university repositories for electronic theses and dissertations [24], Thesis Canada [25], LinkedIn and Proquest [26] to identify individual researchers. Different software were used MS Excel, to find the correlation coefficient and Minitab17 to find ANOVA table. The crawler retrieved real data which was analyzed to highlight the performance of Canadian researchers in the field of computer science on Researchgaye.net. This required analysis of retrieved data to investigate co-authoring between researchers ((MSc student /Supervisor), (Ph.D student/Supervisor))

and co-authoring locally, nationally and internationally with further metric analysis. A Contribution Determines Sequence (CDS) method of listing co-authors on the byline has been proposed giving weight to different author positions. Owing to the fact that authors names can be written in different forms, the crawler retrieved these names regardless whether they belong to the same author or not. On the publications byline on Reserachgate.net authors come into different forms for example the author's name appears full one time, abbreviated another and even using dots symbol [.....]. Name activation was another challenge when it comes to manual author's name filtration, since no consistent method was used to activate authors names on the byline, which required manual verification. Having more than one profile (some have 3) posed another challenge and in case of duplication, the main profile was considered.

1.5 Research Questions

The research addresses the following questions:

- Can a crawler be developed to crawl the academic social network Researchgate.net which lacks an API?
- What scripts can be written? How efficient are the scripts? What analysis can be done on retrieved data from Researchgate.net?
- What method can be developed to accommodate a co-author's contribution in "Impact Points" on Researchgate.net?

1.6 Research Contributions

Crawlers have a history of development associated to the Internet growth.

- A crawler has been designed and developed for Researchgate.net via a series of algorithms.
- A Contribution Determines Sequence (CDS) method has been developed, a new technique for calculating "Impact points" on Researchgate.net.
- To test the crawling system, Researchgate.net has been crawled and researchers data extracted. Data analyzed to demonstrate Canadian researchers' performance on the academic social network Researchgate.net.

1.7 Research Organization

The remaining sections of this research are organized as follows:

- In Chapter 2, a related work to the area of study has been briefly given.
- In Chapter 3, the crawler script implemented and the crawler's architecture shown.
- In Chapter 4, the analysis of the data retrieved after testing the developed crawler has been presented and researchers' performance evaluated.
- In Chapter 5, a new suggested method and scripts for calculating the author's byline position have been presented, to be applied on Researchgate.net to improve the reading of "Impact Points" and consequently the "RG Score".
- In Chapter 6, the conclusion of this research and the future work have been presented.

1.8 Summary

Crawlers have been used to extract information from the Internet sites. They are different in type, purpose and architecture. Social networks and academic social network used by scholars

were on the rise. Our contributions were a crawler, a Contribution Determines Sequence method and highlighting Canadian researchers' performance on Researchgate.net.

Chapter 2

Related Work

This section gives background information about how crawlers and social networks crawling. We split the related work into two categories crawlers and crawling Social Networks. Babaria et al developed a general crawler (Nutch Software) into focused crawler using a plugin ordinal regression module to the Nuthch code. For layered web graph construction, the researcher used Google API. The layered models learns the link leading to topic pages in the form of regression problem which gives the topic pages link distance when solved. To overcome the web large scale nature the researcher proposed a clustering based on Second Order Cone Programming (SOCP). MapReduce programming model was used in making Nutch code. He showed Ordinal Regression(OR) problem overview and used function for mapping data points. Imput Training data and validation data were used during training. Experiment showed that the crawler was efficient in comparison to other crawler [27]. Batsakis et al addressed focus crawlers design and implementation. A new learning crawler is suggested which is a development of earlier Hidden Markov Model (HMM) crawler and the harvest rate performance criteria was established. Different from classic crawlers, it is capable of learning target pages content, and paths leading to target pages. It is able to distinguish between pages assigning the same value. Experiment

results indicated higher performance of the suggested crawler especially when the topic is not clear [28].

In his thesis H. Bakshi proposed a program allowing researchers to collect any kind of data from social networks easily. Scripts were developed to collect data through Twitter API about events and user location. He built an interface that allows collecting and managing data. He described the data collecting method, Twitter API, the database structure. The method he assumes is helpful in particular to journalists who can easily obtain a list of events. To gauge and compare activity between locations, the rate of Twitter tweets per city was used and calculated. The researcher managed to build an efficient framework for collecting data to detect events and trends from the social network Twitter [29]. Z. Xiao et al devised and used a crawler based on interactive simulation to crawl Facebook, with algorithm capable of obtaining whole friend list of any Facebook use, Metropolis-Hasting Random Walk with Delayed Acceptance. Restriction imposed on Facebook pages required the use of the mentioned algorithm. A real user credentials were used to to login in to Facebook, then the number of freiends is extracted and calculated. Crawled datasets showed improved privacy protection among users with higher awareness of females by 16.8%, compared with male users [30]. M. Islam proposed a method to be used on Twitter for recommending new followers to the users, then the best recommendations strategies are identified by an algorithm. He proposed a method using history data to find out the applied strategies the user followed previously. Data crawler,processor and recommender system were designed and implemented to be used in Twitter for following. Data was collected on 3 months time span using a followee-followee ranking Pseudo-code. The researcher reported using machine learning techniques which gave better performance than applying multiple recommendations strategies and this method is applicable to any other social network [31]. S.I. Mfenyana et al reported on a tool with visualization element and frequency analysis module to be used for

collecting data from Facebook, indexing and analysing it. It is possible to query the extracted information to identify what opinions and comments are made about certain topics. There is an extension to the tool to supply and collect words associated with services. The researcher gave an overall view to the way the crawler performs. To reach a certain page, the crawler starts with URL then accesses other pages. Population of the seed URLs list of non-visited nodes occurs by a user or program, followed by a crawling loop. Iterated Incremental approach was used and a system was developed using open source technology in different modules, which were tested continually then integrated into a package. Additional informal interviews to collect data to be familiarized with community people was carried on. The package served to extract data, text-pre-processing and text indexing from Facebook. One user interface included content matching and frequency analysis sub-modules. Crawling Facebook included a back-end java based module on top of alternative to Facebook Graph API called RestFB and Text indexing module was built on top Lucene with another submodule allows keyword searches [32]. O. Almousa and A. Bin Ghazi studied Academia.edu as an academic Social Network, exploring the usage patterns by different researchers. The research importance comes from the fact that there are limited studies on academic social networks in general and on usage patterns in particular. Data was collected and analyzed of researchers from four disciplines Anthropology, Chemistry, Computer Science and Philosophy. Two research questions were put to be answered .One if researchers from different disciplines use Academia.edu differently and the second if academic position has any effect on their use. The study indicated that for Profile completeness faculty members and post-doctoral researchers have top scores. Independent researchers have the least ranking regarding “Relationship” but scored high for “Activity Frequency”. Post-doctoral researchers showed distinct activity in asking and answering questions and highly active regardless of discipline. Computer Science researcher showed higher activity than Chemistry ones [33]. A. Kadriu

studied collaboration networks inside Researchgate. He tried to find out collaboration between academic staff at SEE University as shown on Researchgate, based on their interests. The researcher investigated automatic clustering of researchers grouping based on their relationship, applying four centrality measures :

Degree of Centrality : measuring the most important vertices within the graph.

Closeness Centrality : measuring how long does it take to transfer information from particular node to all other nodes sequentially. The farness of this particular node is defined as the sum of its distance to all other nodes.

Betweenness Centrality : the number of times this node acts as a bridge along the shortest path between other nodes.

Page Rank : measuring the number and quality of links to particular page and this is how the importance of page comes from.

The top topics of interest identified are Computer Science, Computational Intelligence Artificial Intelligence, Economics, Applied Linguistics and Educational Research. The academics who are engaged in different topics of attractions to other a certain faculty member were identified as an influential. Academic Social Network can provide information on research groups in addition to the main goal of collaboration and knowledge sharing. The researcher believes that the same approach can be applied for future research groups such as reviewers and potential MSc./Ph.D. students for example [34]. Z.Tom et al designed and experimented an Academic SNW consisting of new metrics for author and institution ranking. There are different metrics used to show the author Impact Factor depending on criteria such as citations and publication impact factors. Statiscs can easily be obtained from services such as ISI the Web of Knowledge and scopus. The researchers developed a crawler-parser for information extraction in addition to manual work. The study moved beyond traditional metric methods

to authors influence, connections and exposure, by designing different social networks metrics, to compare authors, using similarity study, undertake case studies and open it to other researchers to ensure usefulness. The conclusion they draw is that the designed metrics gave different rankings of authors and allow for better reading of an author. Data was collected from the API of Microsoft Libra. The known ranking metrics focus is on papers, authors and venues captured by the graphs as follow: Papper citation network : Gp=(Vp,Ep). Set of papers is represented with Vp. Citations from one paper to another is represted with Ep. For authorship network GAp=(VA [VP , EAP), where VA is a set of authors, and EAP link of author to paper. For Venueship network GVP =(VV[VP,EVP). The researcher claims more reliable metrics than traditional ones [35]. M. Thelwall investigated the researchers' (Philosophy Scholars) attributes on Academia.edu academic social network. To adjust for time delay in of researchers in making use of Academia.edu a median-based time-normalising was used. The other side to be investigated is if academic impact statistics can contribute to impact estimation. There are limited studies about academic social networks and nearly nothing on the impact of academia.edu in changing scholarly communications. There have been studies on finding metrics through counting tweets citation on Twitter as early indicator of publication impact, besides other altmetric studies that investigated such indicators. The study addressed the questions whether students profiles attract more views, and whether females attract more views than males than males. An other questions are whether senior academics profiles attact more views than others and the type of contents are associated with high profile views. SocSciBot crawler was used on 28 January 2013 to crawl academia.edu at low speed supplemented by a software Webometric Analyst software. www.lexiurl.wlv.ac.uk with use guidance fromwww.lexiurl.wlv.ac.uk/examples/HowToExtractAcademiaI-nfoAboutSubjects.pdf . The results showed that students showed listed slightly more interests than faculty but faculty showed

more publication and views [36].

2.1 Summary

Related work in chapter 2 is divided into two categories crawlers, crawling social networks giving a background information.

Chapter 3

Crawling Researchgate

A crawler system has been created and implemented in Java to crawl data on researchers on Researchgate.net. The system is hosted on our PC and has access to all ours core modules. The crawler is with one thread, uses "Breadth First" searching when searching a repository, using moderate CPU during the search. The crawling is done in order in which it encounters FIFO queue. Before discussing crawler scripts, an overview of the crawler was given, the main features and the Software needed, crawler architecture and the main algorithms built to fetch data from Researchgate.net.

3.1 Crawler (Overview)

3.1.1 Queues :

When a crawler is run on social networks, all web pages, URLs or profiles seeds, need to be pushed in queue in order to be visited. At the beginning, the pages will be addressed as non-visited, but later on, the crawler will update the list(queue), while processing by adding new URLs to be visited.

3.1.2 Crawling Loop :

Crawling loop is an important process, when initializing the web crawler it starts with the first URL in the queue discipline (FIFO), to crawl and fetch information from pages or profiles. In the case of the social network, the crawler will use the URL or user ID for grabbing the contents using hypertext transfer protocol (HTTP), and copy the hypertexts inside these pages for the next visit as well. If the crawler discovers new(non-visited) URLs or users ID, it will be added it to queue(Q) [37].

1. Push the seeds of profiles/URLs into queue and unvisited pages need to be fetched.
2. Check the queue whether it is empty and if not, move to the next URL otherwise end crawling.
3. Fetch the page and grab the contents via Hypertext Transfer Protocol (HTTP).
4. Parse and index data from the page to move to the next step of adding the new URL into queue. Using URL Normalization, the crawler can avoid crawling the same resources more than once.

3.2 Crawler's Features

1. The Crawler is able to crawl any source type (web, databases, File system, CMS).
2. It enables starting any URL, whether a web page, rss feeds or sitemaps.
3. It can be stopped, and resumed crawling.
4. The number of items crawled by source are simultaneous.
5. The crawler can crawl based on item type whether (html, PDF) on periodicity bases rules.
6. Cache crawled items.

3.3 Software Needed

The following software were required to built our crawler:

1. Windows OS.

2. Cygwin tool : is used as a Unix-like environment and command-line interface for Microsoft Windows. Windows applications have been run from the Cygwin environment and used Cygwin tools applications within the Windows operating context providing native integration.

3. Java 6 or 7 : Oracle JDK 6 or 7 have to be installed on the server.

4. MongoDB : Starting version 4.0.0, MongoDB is a mandatory pre-requisite for the following reasons: To store web sites crawl settings, keeping crawled items history, manage crawl queues and managing crawl cache. The database allowed us to write complex queries and to retrieve data about researchers. MongoDB is self-contained and does not have any other system dependencies which can be run from any chosen folder and can be installed in any directory.

5. Apache 2.2.x and PHP 5.3.x/5.4.x : have to be installed on the server.

6. Download (Crawl Anywhere) as a framework to run our scripts.

3.4 A crawler Architecture

The crawler platform consists of MongoDB, admin interface built upon PHP, Apache server and JDK environment. At the begining the Httpclient was an the interface used for communication between crawler programs and researchgate.net, where the crawler program consist of Researchers Crawler, Publication Crawler and Question- Answer Crawler. A Researcher Crawler makes an http client request to Researchgate.net website. Once the user is authorized by Researchgate.net, Researcher Crawler crawls the researcher pages, and finds researcher details saved into the output directory as a csv (comma separated values) file Researchergate.csv.

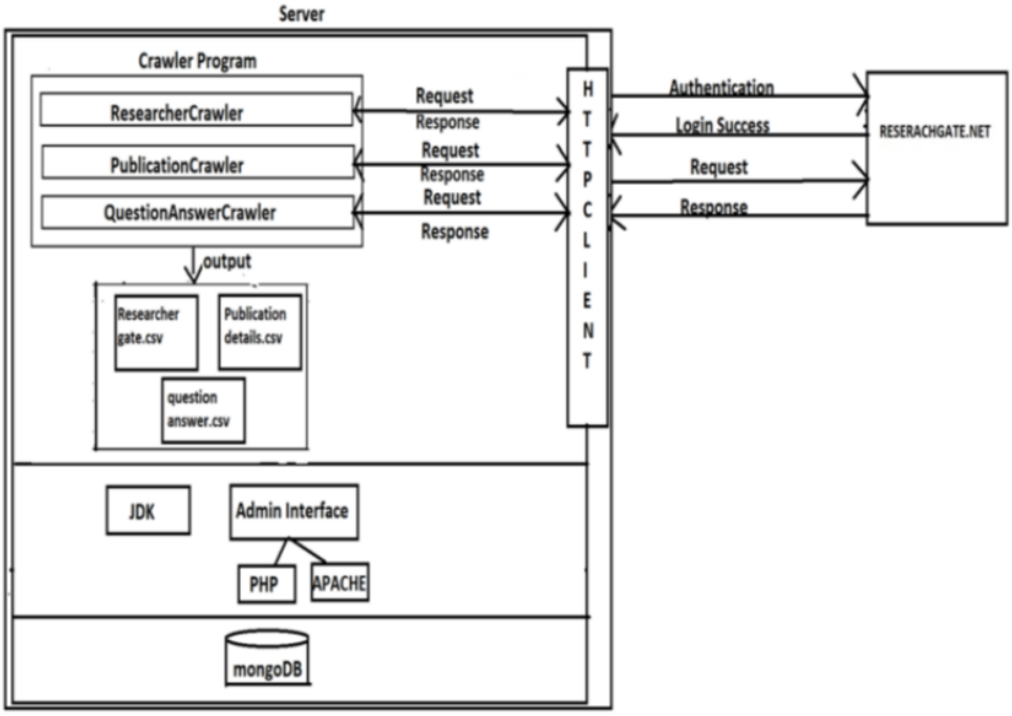


Figure 3.2: Researchgate.net Crawler Architecture.

For the Publication Crawler also the crawler will make an http client request to Researchgate.net website. In case of authorization, the Publication Crawler crawls the publication details of Researchers and saves data into the output directory as a csv (comma separated values) file Publicationdetails.csv. Question and Answer Crawler makes an http client request to Researchgate.net website. When the user is authorized by Researchgate.net, Question and Answer Crawler starts crawling the Question/Answer details of a Researcher and saves data into the output directory as a csv (comma separated values) file Question and Answer.csv.

3.5 Main Algorithms

The three main algorithms built to crawl Researchgate.net to retrieve the following information:

3.5.1 Crawling Researchers' Information

In the absence of API of Researchgate.net, writing algorithms to retrieve researchers' data was important. These algorithms were the basis on which the crawler code was written, thereby making it possible for all the data to be retrieved successfully. Algorithms were built and executed using http client protocol, then URL to access Researchgate.net was prepared. When the request was sent, a response was retrieved back asking for authentication to enter username and password associated with any institute. A ClientLogInForm was prepared visit (Algorithm 2 in the Appendix) for entering the username and password. The request was built and sent, then the response was received for entry, and the first url was built to start fetching data. The url held the university name and other data. A list of 70 Canadian universities and institutes was prepared and saved in a text file to be looked at on Researchgate.net. The file was called from the code by creating file constructor and set the path to reach this text file. The data in this file has been saved calling java.io.BufferedReader class which reads text of character input stream buffering characters to provide an efficient reading of them as an arrays and lines visit (Algorithm 4 in the Appendix). The crawler retrieved data randomly about the universities and institutes, but our (Stop List) was limited to 70. The data in the bufferReader() had to be matched with the data obtained from the crawler and saved inside jsonobject (JavaScript Object Notation) which is a lightweight data-interchange format to make it easy for humans to read and write and for machines to parse and generate. Since several of Canadian institutes have incomplete profiles and some of the Canadian universities with only an account opened on Researchgate.net with no data, the list of (70) universities and institutes, went down to (32) Canadian universities only (Institutes were disregarded in this study). Data was filtered as university id, university name, key university name and the matched data has been saved into an interface called map, which assign unique with no duplicate allowed keys to values. These

data separated into three arrays, since the data captured needed to be rearranged for better understanding and analysis, arrays were important to introduce and store the data, in order of precedence. For each array the whole map has to be assigned to grab data from (ex. we assigned the map to the universityid array to grab the university id and the same for universityname and keyuniversityname) visit (Algorithm 3 in the Appendix).

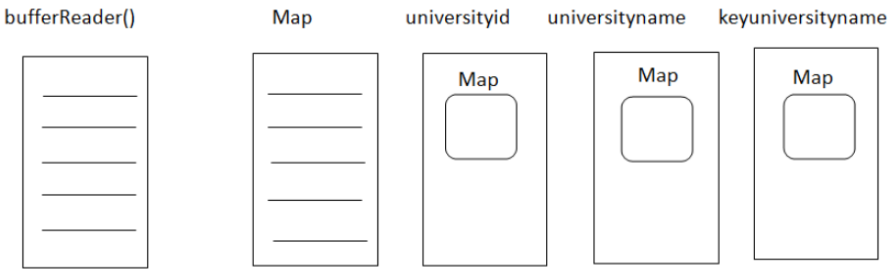


Figure 3.3: A clarification of how to match the data.

Then fetched department of Computer Science for each particular url was required, but since each university has different departments, all these departments were retrieved, and saved into the array called "departname", for each url of department, fetching the Computer Science Department was only needed, so each line saved as a list of string and a counter was used to look at the names "Computer" and "Science", and saved into a separate array "departmentNames", therefore nested arrays have been created visit (Algorithm 1,5 in the Appendix).

After retrieving the data of all Computer Science of the Canadian universities, and having them added into the map, researchers' data retrieved with information like institution, where researcher is being retrieved from (Name of Researcher, Institute, Department, Followers, Publications, Views, Downloads, Citation, Impact Points, RG Score, No. of Questions, No of Answers). At the beginning Computer Science departments array had been picked from the map, researchers crawled within each department and their names have been saved in another array. For Loop used to read the name of the researcher from this array and retrieved the source

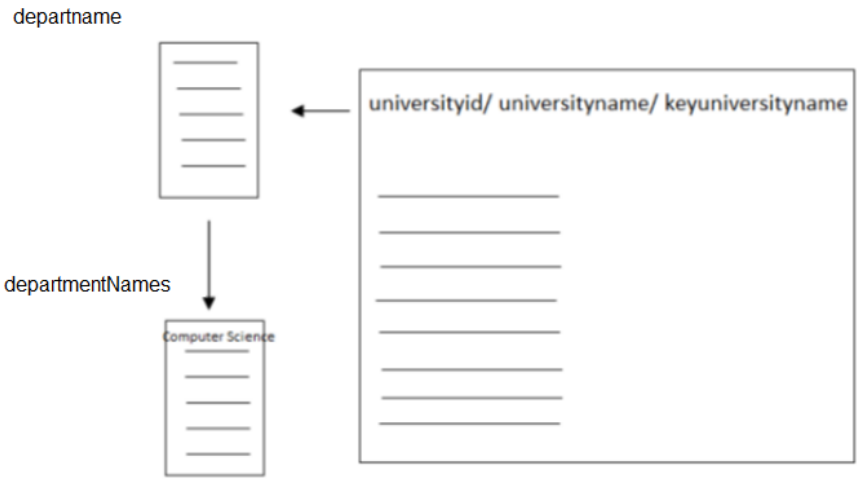
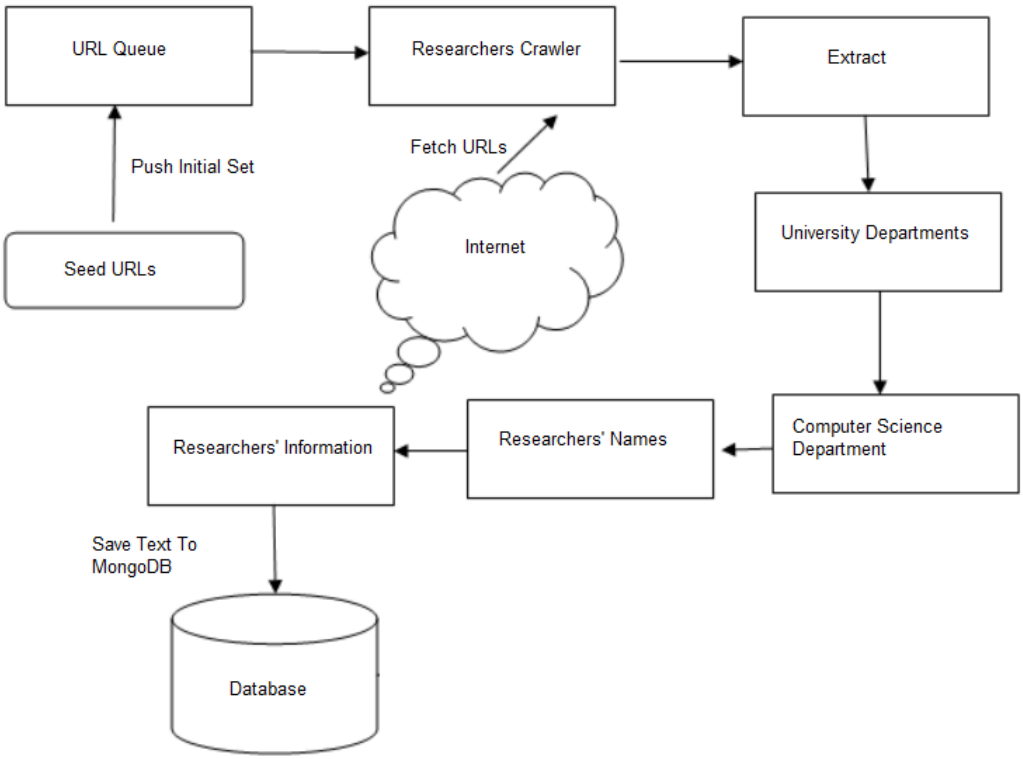


Figure 3.4: A clarification of Fetching Departments of Computer Science.

of his/her document. Researcher array tested whether it was empty or not, if it was all details had to be saved inside it ending the loop and saving the data as .csv file, visit (Algorithm 7 in the Appendix). Jaxen is an open source XPath library used in this crawler to treat non-XML trees such as compiled Java byte code as XML, which make it possible to query trees with XPath. XPath as can be visited in (Algorithm 6 in the Appendix) class given an XML document and tasked with extracting text for known elements. Jericho library used to extract pieces of text from specific locations in the HTML.using Jericho API and Jericho HTML Parser library allowing analysis and manipulation of parts of an HTML document. The following Figure 3.5 shows Researchers' crawler process and data obtained as .csv format:



	A	B	C	D	E	F	G	H	I	J	K	L
1	No.	Name	Institute	Followers	Publications	Views	Downloads	Citations	ImpactPoints	RG Score	No. of Questions	No. of Answers
2	1	Fahiem Bacchus	University of Toronto	28	66	3k	402	1,839	0.2	1.66	0	0
3	2	Ron Baecker	University of Toronto	10	7	486	29	218	0		0	0
4	4	Garry Beirne	University of Toronto	3	3	64	0	60	0	0	0	0
5	5	Patrick Bennett	University of Toronto	7	7	202	28	10	0.35	0	0	0
6	6	Xiaojun Bi	University of Toronto	5	9	373	18	60	0		0	0
7	7	Allan Borodin	University of Toronto	24	21	2k	435	1,696	3.26	5.37	0	0
8	8	Craig Boutilier	University of Toronto	207	213	1k0	3,114	7,305	6.39	9.18	0	0
9	9	Darius Braziunas	University of Toronto	16	8	414	1,887	125	0		0	0
10	10	Julian Brooke	University of Toronto	5	10	406	29	33	0		0	0
11	11	Mathew Brown	University of Toronto	45	16	1k	179	120	17.26	11.57	0	0
12	12	Roy Bryant	University of Toronto	5	2	211	30	6	0		0	0
13	15	Christina Christod	University of Toronto	5	1	48	34	0	0		0	0
14	16	Stephen Cook	University of Toronto	6	11	853	148	58	0.28	1.78	0	0
15	18	Eyal de Lara	University of Toronto	15	101	6k	276	1,466	18.24	17.16	0	0
16	19	Carrie Demmans	University of Toronto	3	13	201	94	7	0		0	0
17	20	Sven J. Dickinson	University of Toronto	59	118	7k	1,817	2,140	66.82	25.87	0	0
18	21	Jiang Du	University of Toronto	3	1	55	33	0	0		0	0
19	23	Ady Ecker	University of Toronto	2	4	353	0	18	0	1.16	0	0
20	24	Faith Ellen	University of Toronto	16	88	4k	579	902	5.68	9.47	0	0
21	25	W.H. Enright	University of Toronto	34	84	3k	490	1,027	58.55	26.6	0	0
22	26	Maryam Fazel-Zar	University of Toronto	10	10	761	275	4	1.08	1.16	0	0
23	27	David J. Fleet	University of Toronto	7	3	159	105	37	0		0	0
24	28	kathleen C Fraser	University of Toronto	2	1	57	0	0	6.16	4.4	0	0
25	29	Stephen Fung	University of Toronto	2	3	244	19	139	0		0	0
26	30	Inmar E. Givoni	University of Toronto	4	16	1k	84	293	36.94	11.01	0	0
27	31	Michael Glueck	University of Toronto	5	12	482	85	24	0.41	1.96	0	0
28	32	Ashvin Goel	University of Toronto	18	33	2k	125	811	0.8	3.74	0	0
29	34	Alex Graves	University of Toronto	32	31	3k	1,187	380	14.93	11.87	0	0
30	35	Joshua A. Grochov	University of Toronto	7	11	585	47	102	11.34	6.22	0	0
31	36	Torsten Hahmann	University of Toronto	15	14	786	3	27	3.91	5.36	0	0

Figure 3.5: Researchers Crawler Process and Data Obtained as .csv Format.

Before starting listing the crawler algorithms details, the following Figure 3.6 shows a Heritage Algorithms Tree.

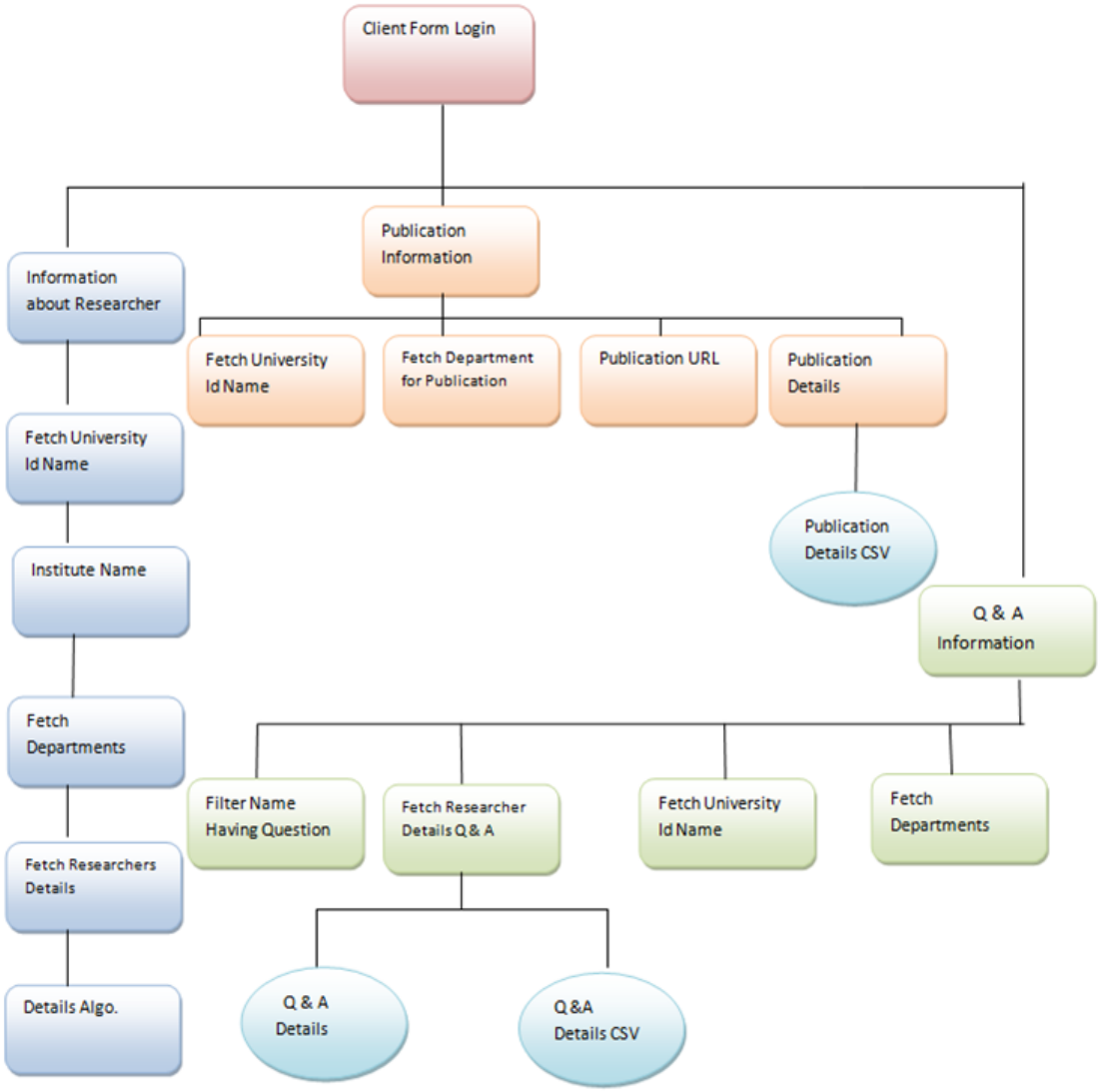


Figure 3.6: A heritage Algorithms Tree.

To facilitate understanding the steps of each algorithm the above Figure 3.6 shows the heritage tree of the three main algorithms used to crawl Researchgate.net and the sub main ones. For further details about the pseudo code of each one visit (Appendix). The codes have

been moved there, while to understand the task of each one of them the first seven algorithms used to crawl information about researchers as follow:

Algorithm 1: Allowed the crawler to identify parameters needed to enable accessing Researchgate.net website, and identifying the possibility to use username and password to login to Researchgate.net. For more details visit (Algorithm 1 in the Appendix).

Algorithm 2: This part of the code used HTTP-GET to request data or information from Researchgate.net servers and client by logging in with Researchgate.net username and password, using POST to submit the data required to be processed. Both the HTTP GET and POST were methods used to send and retrieve information from Researchgate.net. For more details visit (Algorithm 2 in the Appendix).

Algorithm 3: This algorithm allowed writing a query and returned all institutions and universities in Canada. For more details visit (Algorithm 3 in the Appendix).

Algorithm 4: This algorithm was written to filter through all the universities and institutions returned in algorithm 3 above to return only Canadian universities/Institutions. For more details visit (Algorithm 4 in the Appendix).

Algorithm 5: Since information on researchers in computer science departments and institutions was sought, this algorithm was written to return data on researchers in the departments of Computer Science at Canadian institutions or universities, and this information was returned accurately. For more details visit (Algorithm 5 in the Appendix).

Algorithm 6: Here in this algorithm again the parameters specified, which were to be returned with the query to return some specific information on researchers from computer science departments at Canadian institutions and universities. Some of the returned data was a total number of impact points, total number of times a researchers publications were downloaded, departments, citations, etc. For more details visit (Algorithm 6 in the Appendix).

Algorithm 7: In this algorithm all the details mentioned in the previous algorithm were collected, but there was a need to specify a file format and a location to append all these details so retrieved data was appended into a specified .csv format in this algorithm to allow data statistically calculated in excel for accurate results. For more details visit (Algorithm 7 in the Appendix).

3.5.2 Crawling Publications Information

After setting the basic cookies store visit (Algorithm 8 in the Appendix), the same algorithms were used in crawling information about researchers used here for getting and saving all the Computer Science departments in the departmentnames array visit (Algorithm 3 and Algorithm 5 in the Appendix). Publications URLs had to be fetched from this array, and these URLs inside array called publication names were saved. Each URL in a list of string called publicationurl was saved and details of each of these URLs had to be crawled using publicationdetails constructor. Details of the publication such as S.N, Publication Name, Institute, Views, Downloads, Date of
Publishing, Main Author, Co-Author1, Co-Author2,....., Co-Author20 had been crawled. In order to avoid returning the same URL twice inside publication names array, certain methods were used to test whether the array had the same publicationurl or not visit (Algorithm 9 in the Appendix). After preparing the array list, constructor publicationdetails was called and a new array list had been created, each URL contained the name of publication. There would be no need for the whole URL, but only the name (title of publication) was required.
It was quite clear that any publication posted on Researchgate.net appeared in the URL box with the title of publication. There was a sub header located at the left of this title, certain classes had to be used in order to separate this sub header in a list of elements this list would have id starting from (0) to be generated from the code. Another class was used to retrieve the

contents of Authors names on their byline positions at position(0) , these contents have been saved into authorsNames array. After that the date of publishing had to be fetched, It was clear that the date of publishing appeared on Researchgate.net right away down the authors names, and the date usually ended with (;) followed by certain numbers. The date before (;) was separated, and tested whether existed or not, or the date had no ("/"). Views and Downloads were retrieved and tested whether existed or not, if yes they have been added into the array, and finally the whole contents of publication details added into a biggest container (constructor) was called publicationdetails visit (Algorithm 10 in the Appendix).

The following algorithms for crawling information about publications :

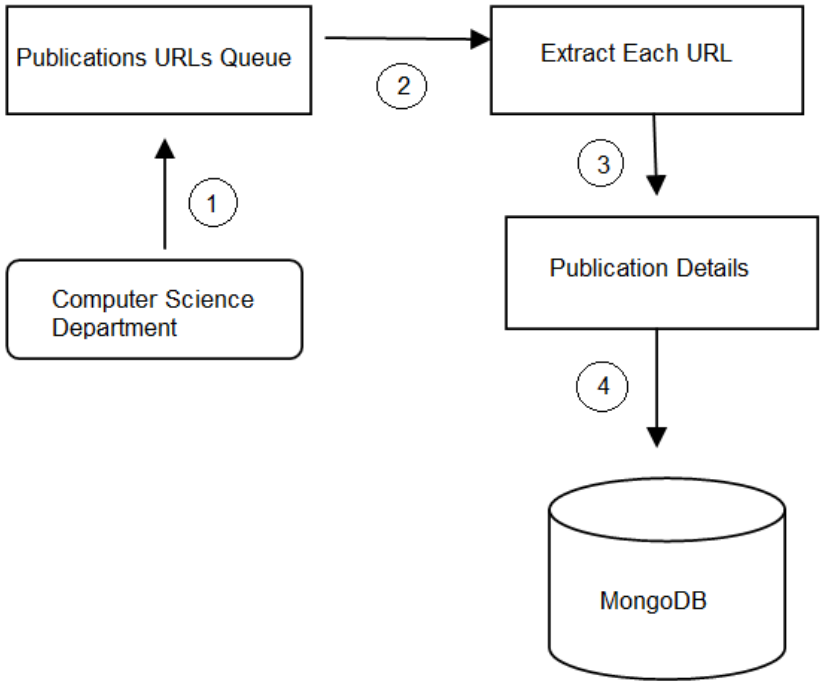
Algorithm 8: This particular algorithm was written to get data on all researchers in Canadian Institutions and universities who had published papers with details on their publications. Here, a complex algorithm was written due to the nature of Researchgate.net by which they indexed their data, and results were attained. Since publication details on researchers from computer science were needed (Algorithm 3 and Algorithm 5) were used here as well, helping provide a filter for algorithm 8 to get more accurate results and after the filtration, all publication details in Computer Science departments from Canadian universities or institutions were retrieved. For more details visit (Algorithm 3, Algorithm 5 and Algorithm 8).

Algorithm 9: A publication class was specified for this algorithm to enable retrieving publication URL based on country, state university and department where researcher's work was published. For more details visit (Algorithm 9 in the Appendix).

Algorithm 10: This algorithm also used GET method to return the requested data in algorithm 9 by sorting through indexes of Researchgate data to get the data run by the query provided in the request. For more details visit (Algorithm 10 in the Appendix).

To append the data in to .csv file visit (Algorithm 11 in the Appendix), the path and file format

have been set to append the information retrieved on publications in to the .csv format, thereby specifying each cell and column to allocate each information retrieved from Researchgate. Different methods used to append these data and others used to write the data inside each cell and to move to new line after filling the information record of each publication. The comma used in the code has been represented as a column in MS Excel, (institute name ,views) were appended. Views was checked whether it was =null or not. If yes, it would be set it to empty in the .csv file and the same for the (Downloads, Date of Publishing). Regarding the names of Authors on their byline position were saved in an array, the size of this array was (20), because most of the papers read were not more than (20) authors. Variable was declared to start reading and saving each picked name from the list in to .csv file. The process was continued till the whole information was completely appended successfully.



	S.NO.	Publication Name	Institute Name	Views	Downloads	Date Of Publication	Main Author	Co-Author1	Co-Author2	Co-Author3	Co-Author4	Co-Author5	Co-Author6	Co-Author7	Co-Author8	Co-Author9	Co-Author1
2	1	Learning structural e	University of Tor	2		12-Jan	J. Chua	I. Givoni	R. Adams	B. Frey							
3	2	Distributed Ranked D	University of Tor	68		13-Jan	Kaiwen Zhang	M. Sadoghi	V. Muthusam	H.-A. Jacobsen							
4	3	Identifying mRNA se	University of Tor	76	1	14-Mar	Jingjing Li	Taehyung Kim	Razvan Nutiu	Debashish Ray	Timothy R Hı	Zhaolei Zhang					
5	4	Next Generation Dat	University of Tor	91	0	Aug-00	Susan Elliott Sim										
6	5	Applying Machine Le	University of Tor	104	0	Jul-99	Susan Elliott Sir	Prof G. Karakoulas									
7	6	The Coming of Softw	University of Tor	69	0	Feb-99	Susan Sim										
8	7	Workshop on standa	University of Tor	130	9	Feb-00	S.E. Sim	R. Holt	R. Koschke								
9	8	Using benchmarking	University of Tor	79		3-Jun	S.E. Sim	S. Easterbrook	R.C. Holt								
10	9	The ramp-up probler	University of Tor	73	0	May-98	S.E. Sim	R.C. Holt									
11	10	Word Segmentation	Ryerson Universit	39	22	Jan-99	Chi-Hung Chi	Chen Ding	Andrew Lim								
12	11	User-centered desig	Ryerson Universit	37			Delnavaz Mobe	Chen Ding									
13	12	Donor insemination:	Ryerson Universit	58		May-96	C L Wendland	F Burn	C Hill								
14	13	The compromise wit	Federal Universit	70	0	6-Oct	S.G.C. Fraiha	J.C. Rodrigues	H.S. Gomes	G.H.S. Carvalho	G.P. Cavalcante						
15	14	Performance analysi	Ryerson Universit	57	0	8-Jan	Glaucio H.S. Ca	Victor S. Martins	Carlos R.L. Fı	João C.W.A. Co	Solon V. Carvalho						
16	15	Optimal call admissi	Ryerson Universit	52	2		G. H. S. Carvalh	C.R.L. Frances	S. V Carvalhc	R. C. M. Rodrigues							
17	16	Analysis of traffic m	Ryerson Universit	66		6-Oct	G. H. S. Carvalh	R.S.C. Cruz	V.S. Martins	C.R.L. Frances	J.C.W. Costa	S. V Carvalho					
18	17	A semi-Markov decis	Ryerson Universit	46		13-Jan	Glaucio H.S. Ca	Isaac Woungang	Alagan Anpal	Rodolfo W.L. C	João C.W.A. Costa						
19	18	SH 2010: Welcome n	Ryerson Universit	27		10-Jan	Jh Park	Lt Yang	S Zeadally	I Woungang	E Law	A Ferworn	A Anpalagan	F Kawsar	Y -S Jeong	E Cerqueira	S Fowler
20	19	SH 2010: Welcome n	Ryerson Universit	23		10-Jan	JH Park	LT Yang	S Zeadally	I Woungang	E Law	A Ferworn	A Anpalagan	F Kawsar	Y-S Jeong	E Cerqueira	others
21	20	Modelling and Perfo	Ryerson Universit	64	3	4-Jan	Glaucio H. S. Ce	R. M. Rodrigues	Carlos Renat	João Crisóstom	Solon V. Carvalho						
22	21	A Markovian Sensibil	Federal Universit	51	2	6-Jan	Regiane Y. S. Ka	Luiz Affonso Guede	Diego L. Carc	Carlos Renato L	Glaucio H. S.	Solon V. Carva	João Crisósto	Marcelino S. da Silva			
23	22	Um modelo de dese	Federal Universit	71	0		Regiane Y Kawa	Luiz A Guedes	Carlos R L Fr	João C W A Cos	Glaucio H S (	Diego L Cardo	Marcelino S S	Luiz D C Augusto			
24	23	Política ótima para c	Ryerson Universit	39			G. H. S. Carvalh	R. C. M. Rodrigues	S. V Carvalhc	FRANCES	C. R. L	COSTA	J. C. W. A				
25	24	Análise de Desempe	Ryerson Universit	20	0	2-Jan	G. H. S. Carvalh	J. C. W. A. Costa									
26	25	Design of optimal Ca	Ryerson Universit	91	69	9-Dec	G.H.S. Carvalho	R.W.L. Coutinho	J.C.W.A. Costa								
27	26	An iterative, multi-le	Ryerson Universit	19			Andriy V. Miran	Nazim H. Madhavji	Mechelle S. (	Matthew Davisc	Mark Wilding	David Godwin					
28	27	Characteristics of m	Ryerson Universit	106		11-Jan	Zude Li	Nazim H. Madhavji	Syed Shariya	Mechelle Gitten	Andriy V. Mir	David Godwin	Enzo Cialini				
29	28	Factors characterizir	Ryerson Universit	18		12-Sep	Bora Cau glaya	Ayc Tosun M 1s	Andriy Miran	Burak Turhan	Ayc Bener						
30	29	Sensor Data Assimila	Ryerson Universit	60	17		Marcus V Santo	Paulo E Santos	P Ac Santos@	Uk							
31	30	A PATH SEMANTICS	Ryerson Universit	34	6		Marcus V Santo	Paulo E Santos									
32	31	Adaptive Represente	Ryerson Universit	49	7	10-Jan	Nigel P. A. Brow	Marcus V. dos Santos									

Figure 3.7: Crawling Publication Details and Data Obtained as a .csv Format.

3.5.3 Crawling Information on Questions & Answers

BasicCookiesStore was created visit (Algorithm 12 in the Appendix), to store data and build httpclient,then logged into the Researchgate.net using username and password, and got the httpclient. Institute id,name,key were fetched, the previous algorithm visit (Algorithm 5 in the Appendix) was used and, the name of the institute was fetched and saved in the array names. From the array keyinstituteName departments array was created, then the whole map passed it to it, to get keyinstituteName. After having the list of keyinstituteNames inside departments array, each URL was crawled and tested whether empty or not if not, that meant the URL existed and there were contents. Different methods were used to get the content and save it in another format to make it easy to read and write and easy for the machine to process it.

(Algorithm 12 in the Appendix) dealt with the interaction between researchers on Research-gate. The way researchers asked questions was picked, and who interacted with them by answering their questions and establishing a relationship with them. The names with ques-tions were filtered, then researchers with questions and answers details were fetched such as researcher's name, and affiliation. Fetching questions details was done, such as the title of the question, the number of answers for that question, with more details about direct answers to the researcher's question. Finally data was added to the .csv file. Profile source code was acquired of a researcher and xpath was used for checking question visit (Algorithm 13 in the Appendix). If there were questions, they were added to the arraylist, else searching for the next researchers was done, and an array named (Names) containing the names of all researchers retrieved from Authornames arraylist was called, for this part of coding, the previous algorithm visit (Algo-rithm 3 in the Appendix) was used. Link to the researchers' contribution was prepared and assigned to the profileURLprefix visit (Algorithm 14 in the Appendix), request sent, the whole

link was static except the name of researcher was dynamic. “For Loop” was used to read researcher’s name then, the profileurlprefix had to be changed to profileurlprefix = profileurlprefix + name. The profileurlprefix tested whether =null or not. If not, the URL was set and source was obtained. Request builder was requested and built, link executed using httpclient, entity of the data was acquired. Methods used to get the data, Arraylist was created to save the source of researchers’ profile. In order to acquire questions only, certain classes were used to fetch each line by it’s index. Declared variables were used to store the question by calling class, for example question at position(0) and so on. The contents of the question’s text extracted and summery tested whether = Questions or not. Once all these steps were done, this result was tested if any, researchers’ name was added into array called researcherName, otherwise it should have shown null. Finally profileurlprefix was replaced by(“ ”) then the step after fetching researcher’s details QA Algorithm was called. Profile source was obtained, and xpath was used for getting the name, institute, department and setting those to bean, then bean object was used as a value and researcher name as a key. (Algorithm 13 in the Appendix), this algorithm filtered the questions and answers results generated by the query to retrieve all questions asked by each individual researcher. To fetch researcher details question/answer,“For Loop” was used to go through researchernames array, and start preparing profileurlprefix, instance was created for researcherHavingQuestionDetails. For the first profileurlprefix, a document was acquired and assigned to source. Again this source was evaluated, results obtained and tested to which type of elements was printed. After running switch case to set researcher’s name, university, department, researcher data should be tested, whether it was null or not, and if not, researcher’s name was added in to the map. Finally profileurlprefix was replaced with (“ ”) instead of (/), then question and answer details algorithm was called to save the data in to .csv file.

(Algorithm 14 in the Appendix), this algorithm allowed to locate which path or index a

particular question originated from and sorting them according to researchers from Computer Science departments in Canadian universities/institutions, and it was possible to find which researcher asked a particular question on Researchgate.net.

Algorithm 15: This algorithm allowed us to specifically identify the responses generated to questions asked by researchers, and to locate the relationship between researchers such as which question was answered by what number of followers, and to determine if all who answered the questions are from the same department as the researcher who asking the question. By this algorithm the relationship between researchers from Computer Science Departments with other researchers from other departments was determined. For more details visit (Algorithm 15 in the Appendix).

Algorithm 16: After filtering all the information submitted by the other algorithms, returned data would be appended into the .csv file to help identifying the researcher behind the ques-tion, and the researcher giving answers to allow analysis and determine their relationships. For more details visit (Algorithm 16 in the Appendix). Figure 3.8, shows the crawling process of researchers details asking questions and details of researchers answers questions. Data obtained from this crawler has been saved as .csv format as shown in figure 3.9.

3.6 Classes Used

The main classes used to extract information from Researchgate.net as listed below:

1. Import import net.htmlparser.jericho.Config.

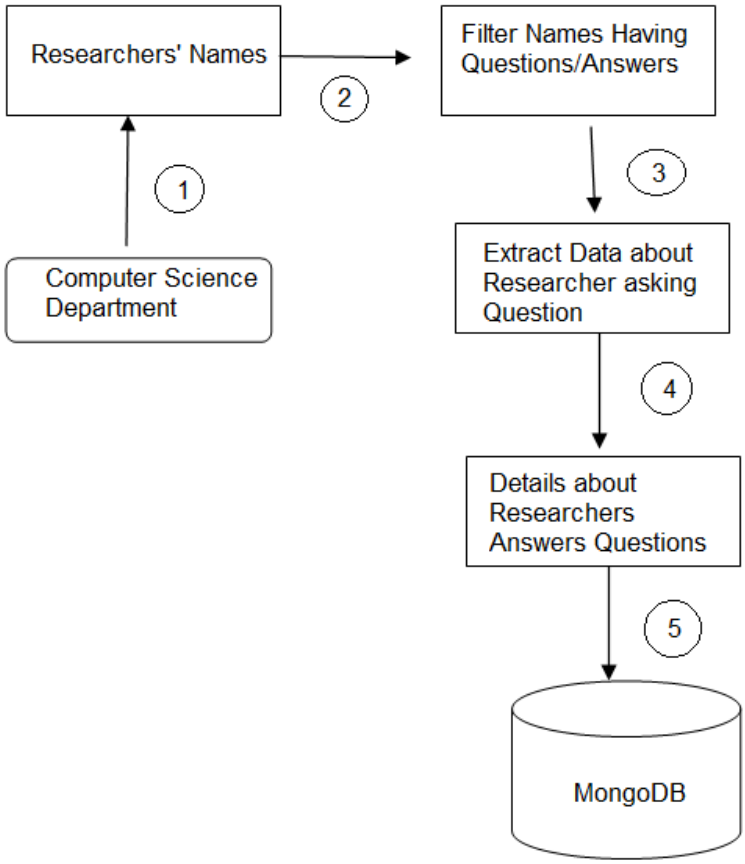


Figure 3.8: Crawling Questions/Answers Details.

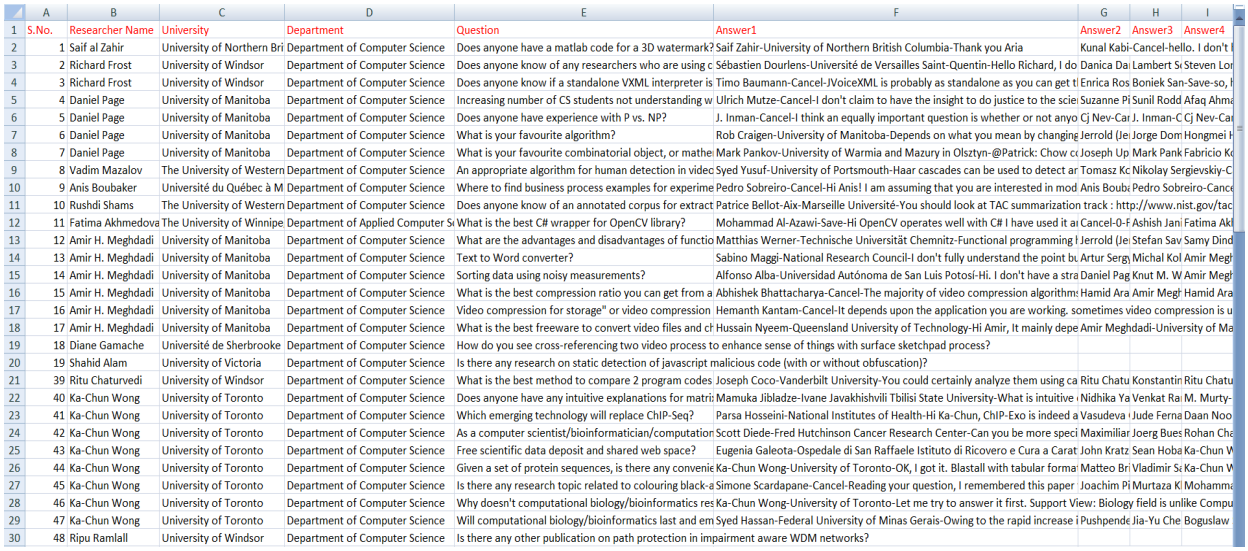
2. Import import net.htmlparser.jericho.LoggerProvider, used to disables all log messages.

3. Import import net.htmlparser.jericho.Source, used for getting source (response) of url.

4. Import import net.htmlparser.jericho.Element, Element class is used for getting elements from source.

5. Import import net.htmlparser.jericho.Attribute, from element we can get attribute value.

An HTTP cookie is a token or short packet of state information that the HTTP agent and the target server can exchange to maintain a session. Some of the important classes used here are:



S.No.	Researcher Name	University	Department	Question	Answer1	Answer2	Answer3	Answer4
1	Saif al Zahir	University of Northern Bri	Department of Computer Science	Does anyone have a matlab code for a 3D watermark?	Saif Zahir-University of Northern British Columbia-Thank you Aria	Kunal Kabi	Cancel-hello. I don't	
2	Richard Frost	University of Windsor	Department of Computer Science	Does anyone know of any researchers who are using c	Sébastien Dourlens-Université de Versailles Saint-Quentin-Hello Richard, I do	Danica Da	Lambert S	Steven Lor
3	Richard Frost	University of Windsor	Department of Computer Science	Does anyone know if a standalone VXML interpreter is	Timo Baumann-Cancel-JVoiceXML is probably as standalone as you can get t	Enrica Ros	Boniek San	Save-so,
4	Daniel Page	University of Manitoba	Department of Computer Science	Increasing number of CS students not understanding w	Ulrich Mutze-Cancel-I don't claim to have the insight to do justice to the scie	Suzanne Pi	Sunil Rodd	Afaq Ahma
5	Daniel Page	University of Manitoba	Department of Computer Science	Does anyone have experience with P vs. NP?	J. Inman-Cancel-I think an equally important question is whether or not anyo	Cj Nev-Car	J. Inman-C	Cj Nev-Car
6	Daniel Page	University of Manitoba	Department of Computer Science	What is your favourite algorithm?	Rob Craigen-University of Manitoba-Depends on what you mean by changing	Jerrold (Je	Jorge Dom	Hongmei H
7	Daniel Page	University of Manitoba	Department of Computer Science	What is your favourite combinatorial object, or mathe	Mark Pankov-University of Warmia and Mazury in Olsztyn-@Patrick: Chow c	Joseph Up	Mark Pank	Fabricio K
8	Vadim Mazalov	The University of Western	Department of Computer Science	An appropriate algorithm for human detection in video	Syed Yusuf-University of Portsmouth-Haar cascades can be used to detect ar	Tomasz Kc	Nikolay Sergievskiy-C	
9	Anis Boubaker	Université du Québec à M	Department of Computer Science	Where to find business process examples for experime	Pedro Sobreiro-Cancel-Hi Anis! I am assuming that you are interested in mod	Anis Boub	Pedro Sobreiro-Cance	
10	Rushdi Shams	The University of Western	Department of Computer Science	Does anyone know of an annotated corpus for extract	Patrice Bellot-Aix-Marseille Université-You should look at TAC summarization track : http://www.nist.gov/tac			
11	Fatima Akhmedova	The University of Winnipe	Department of Applied Computer Sc	What is the best C# wrapper for OpenCV library?	Mohammad Al-Azawi-Save-Hi OpenCV operates well with C# I have used it ar	Cancel-0-F	Ashish Jani	Fatima Akh
12	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	What are the advantages and disadvantages of functio	Matthias Werner-Technische Universität Chemnitz-Functional programming l	Jerrold (Je	Stefan Sav	Samy Dind
13	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	Text to Word converter?	Sabino Maggi-National Research Council-I don't fully understand the point be	Artur Serg	Michal Kol	Amir Megh
14	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	Sorting data using noisy measurements?	Alfonso Alba-Universidad Autónoma de San Luis Potosí-Hi. I don't have a stra	Daniel Pag	Knut M. W	Amir Megh
15	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	What is the best compression ratio you can get from a	Abhishek Bhattacharya-Cancel-The majority of video compression algorithms	Hamid Ara	Amir Megh	Hamid Ara
16	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	Video compression for storage" or video compression	Hemanth Kantam-Cancel-It depends upon the application you are working. sometimes video compression is u			
17	Amir H. Meghdadi	University of Manitoba	Department of Computer Science	What is the best freeware to convert video files and ch	Hussain Nyeem-Queensland University of Technology-Hi Amir, It mainly depe	Amir Megh	dadi-University of Ma	
18	Diane Gamache	Université de Sherbrooke	Department of Computer Science	How do you see cross-referencing two video process to enhance sense of things with surface sketchpad process?				
19	Shahid Alam	University of Victoria	Department of Computer Science	Is there any research on static detection of javascript malicious code (with or without obfuscation)?				
39	Ritu Chaturvedi	University of Windsor	Department of Computer Science	What is the best method to compare 2 program codes	Joseph Coco-Vanderbilt University-You could certainly analyze them using ca	Ritu Chatu	Konstantin	Ritu Chatu
40	Ka-Chun Wong	University of Toronto	Department of Computer Science	Does anyone have any intuitive explanations for matri	Mamuka Jibladze-Ivane Javakhishvili Tbilisi State University-What is intuitive	Nidhika Ya	Venkat Rai	M. Murty-
41	Ka-Chun Wong	University of Toronto	Department of Computer Science	Which emerging technology will replace ChIP-Seq?	Parsa Hosseini-National Institutes of Health-Hi Ka-Chun, ChIP-Exo is indeed a	Vasudeva	Jude Ferna	Daan Noo
42	Ka-Chun Wong	University of Toronto	Department of Computer Science	As a computer scientist/bioinformatician/computation	Scott Diede-Fred Hutchinson Cancer Research Center-Can you be more speci	Maximiliar	Joerg Bues	Rohan Cha
43	Ka-Chun Wong	University of Toronto	Department of Computer Science	Free scientific data deposit and shared web space?	Eugenia Galeota-Ospedale di San Raffaele Istituto di Ricovero e Cura a Carat	John Kratz	Sean Hoba	Ka-Chun W
44	Ka-Chun Wong	University of Toronto	Department of Computer Science	Given a set of protein sequences, is there any convenie	Ka-Chun Wong-University of Toronto-OK, I got it. Blastall with tabular forma	Matteo Bri	Vladimir Sa	Ka-Chun W
45	Ka-Chun Wong	University of Toronto	Department of Computer Science	Is there any research topic related to colouring black-a	Simone Scardapane-Cancel-Reading your question, I remembered this paper	Joachim Pi	Murtaza Kl	Mohamma
46	Ka-Chun Wong	University of Toronto	Department of Computer Science	Why doesn't computational biology/bioinformatics res	Ka-Chun Wong-University of Toronto-Let me try to answer it first. Support View: Biology field is unlike Compu			
47	Ka-Chun Wong	University of Toronto	Department of Computer Science	Will computational biology/bioinformatics last and em	Syed Hassan-Federal University of Minas Gerais-Owing to the rapid increase i	Pushpende	Jia-Yu Che	Boguslaw
48	Ripu Ramlall	University of Windsor	Department of Computer Science	Is there any other publication on path protection in impairment aware WDM networks?				

Figure 3.9: Questions/Answers Data Obtained as .csv Format.

1. import org.apache.http.impl.client.BasicCookieStore;

2. import org.apache.http.impl.client.CloseableHttpClient;

3. import org.apache.http.client.methods.CloseableHttpResponse;

4. import org.apache.http.client.methods.RequestBuilder;

5. import org.apache.http.client.ClientProtocolException;

3.7 Summary

Web crawlers have been used as the main component of web search engines. Since Researchgate lacks an API, a sophisticated scripts crawler have been created and implemented in Java to collect data about researchers on Researchgate.net. An overview of the crawler, the main features and the software needed are discussed in addition to the crawler architecture, algorithms are discussed as well.

Chapter 4

Testing the Crawler

To test the suggested crawler, Real-life data is collected from Researchgate.net. Data was extracted from Canadian researchers in the field of Computer Science and their performance on Researchgate.net using our crawler to retrieve the following data from a researcher's profile:

a -"RG Score" : readings that determine the standing of the individual (Or the department) standing among others on Researchgate.net.

b -"Impact Points": measures an authors' impact factor and all his activities on Researchgate.net giving an early indicator of his publication impact.

c - Publications: An author's publication posted on his page.

d - Downloads: The number of times an author's publication are downloaded by other researchers.

e - Views: The number of clicks done to view author's profile.

f - Followers: Other researchers who follow an individual researcher on Researchgate.net.

This academic social network, allows users to create their profiles, upload their papers, download papers of interest, and join research groups. The data analyzed in this work for a period of three months starts from April 3- June 28 ,2015. A sample collected of (506) researchers and analysis indicated:

4.1 Collaboration-Co-authoring

Collaboration usually occurs among researchers. The extracted data was analyzed to highlight collaboration on Researchgate.net through co-authoring and Question/Answer activities. Metrics is the other analysis perspectives. Co-authored researches are researches where two or more participated to carryout and finalize the research. Researchers were identified to study the co-authoring into four different categories:

1. Supervisors.
2. MSc. students.
3. Ph.D students.
4. Co-supervisors.

4.1.1 Co Authoring between Supervisors and Students (MS.c.)

Since the student is a new researcher, being trained on academic research, they are encouraged to publish jointly with more experienced researchers (their supervisors). A filtered sample of (112) pairs of students/supervisors and the joint publications. Data was collected using Excel Sheet and percentage of joint publications was calculated using the following formula:

$$Percentage of Joint Publications = \frac{JwP}{MsP + SuP - JwP} * 100 \qquad (4.1.1.1)$$

Where:

JwP: Joint publications.

MsP: MSc. student publications.

SuP: supervisor publications.

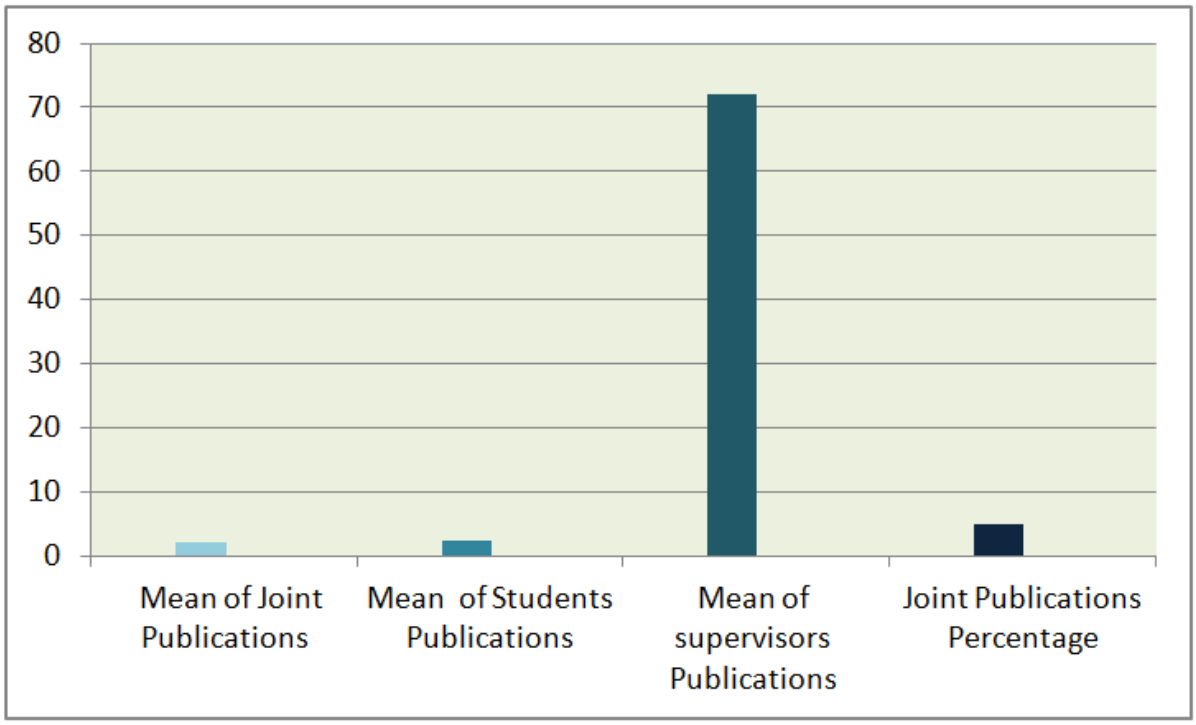


Figure 4.1: Supervisor/Student (MS.c.) co-authoring.

Figure 4.1, indicates that graduate students at this level prefer to publish with reputable supervisors, and this can be attributed to their limited research skills, that require close supervision and guidance from more experienced researchers. Euler and Venn diagrams were used to represent the percentage of joint publications.

Euler and Venn Diagrams

Euler and Venn Diagrams are a set of circles, ovals or ellipses representing a set of relationships by interconnected curves, of which one set can be partial of the other set. Euler curves are divided into two zones, interior representing the elements of the set and exterior representing the elements which are not part of the set. A Venn diagrams can be Euler diagrams, but not all Euler Diagrams can be Venn diagrams. Venn represents all possible relations between sets (zones) and contains all of them, while Euler only contains subset of these zones. The set of circles can be overlapped as well colored. Euler can be a set of two or more while Venn diagram is three and more. With the increasing number of circles the diagram becomes visually complex, while a shaded Zone in Venn represents an empty set, the corresponding zone in Euler is missing from the diagrams. To conclude that the number of contours have been increased, Euler diagrams become less complex than Venn diagram, especially if the number of the non-empty intersections are small. The following graph shows a Euler and Venn diagram.

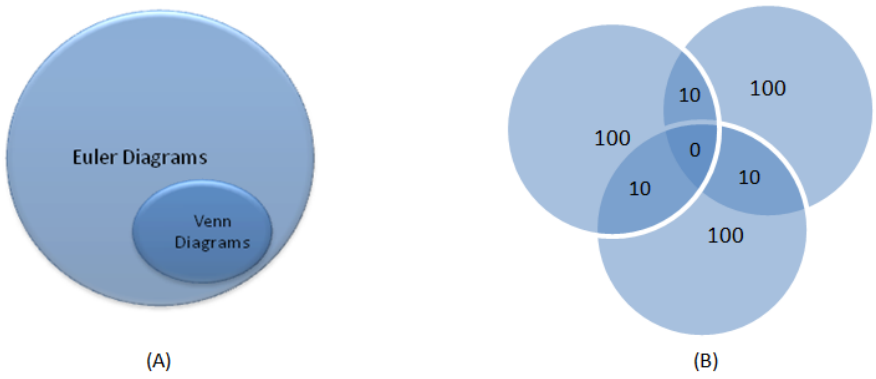


Figure 4.2: Euler and Venn Diagram.

Figure 4.2 (A), shows Euler diagram where Euler contains of number of close curves, which shows how different groups of things are related and some of these curves may be wholly contained in another curve. Figure 4.2 (B) shows Venn Diagram with the non-exist area or the

area represents in zero as a content, while the rest of areas are considered Euler Diagram.

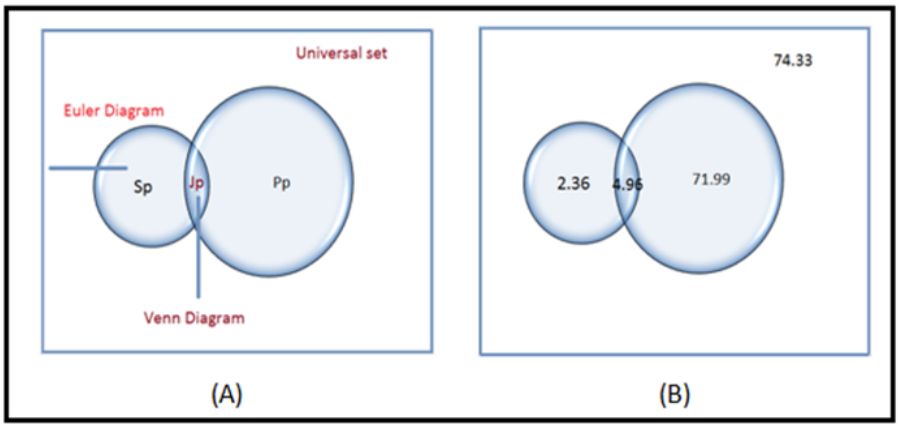


Figure 4.3: Euler and Venn diagrams for joint supervisor/student(MS.c.) publications.

Figure 4.3 (A), shows the relationships between supervisors' publications and MSc. Students represents in symbols. Figure 4.3 (B) shows the mean of supervisor/student publications represented in Euler Diagrams. Shaded area represents the percentage of joint publications between them in Venn Diagrams.

4.1.2 Co-authoring between Supervisors and Students (Ph.D.)

The same approach used to study joint publishing collaboration between supervisors and MSc. students, was applied to study Ph.D supervisor/student joint publications. A filtered sample of (169) in pair supervisor/student(Ph.D) was collected and organized on Excel sheet. Care was taken through data collection to consider when a Ph.D student publishes with a supervisor one time and with a co-supervisor another. Consideration is made also when a supervisor published with a co-supervisor only. The highest percentage of joint publications among Ph.D students/supervisors/co-supervisors was (%7) and Ph.D students with only a principal supervisor was (%35) and it was calculated the percentage of supervisors' total publications to the joint ones. The result was organized in two tables. The first contains the mean of

student/supervisor/co- supervisor publications along with the mean of joint publications among them.

The second table contains the percentage of their publications and the percentage of joint ones. After finding the mean (First Moment(μ)), it was necessary to find the Second Moment (M2) to describe the way that the probability density function distributed about its mean. The variance and standard deviation were considered for this purpose, where the variance describes whether the distribution of our date is clustered close to its mean, or distributed over a great distance from the mean.

Table 4.1: First and second moments along with the variance and Stdev for Ph.D students/supervisors/co supervisors publications.

	Mean(First Moment) E(x) or(μ)	Second Moment E(x2) or M2	Variance (V)	Stdev(σ)
Students publications(Sp)	7.631	81.012	22.78	4.773
Supervisor publications(Pp)	90.548	12776.774	4577.903	67.66
Co- supervisors publications(Cp)	39.733	1980.667	401.956	20.049
Joint students/supervisors publications (Jp)	5.911	49.756	14.819	3.849
Joint students/ co- supervisors publications(Jc)	3.267	13.667	2.994	1.7303
Joint supervisors/co- supervisors publications(JPc)	14.2	303.133	101.493	10.074
Joint /stud./sup./co- sups. publications(JSPc)	1.067	1.733	0.595	0.771
Total supervisor publications(TOp)	97.821	14318.69	4749.742	68.918

Table 4.2: The percentage of actual and joint publications.

	Mean(First Moment) (μ)	Second Moment M2	Variance (V)	Stdev(σ)
The percentage of supervisors publications(ROp)	90.406	8237.006	63.761	7.985
The percentage of co-supervisors publications(ROc)	37.282	1768.813	378.865	19.464
The percentage of students publications(ROs)	11.189	239.701	114.507	10.701
The percentage of joint students/supervisors publications (RJp)	8.354	115.974	46.185	6.796
The percentage of joint students/ co supervisor publications (RJc)	3.526	15.892	3.099	1.76
The percentage of joint supervisors/co -supervisors publications(RJPc)	12.55	198.478	40.963	6.4
The percentage of joint students/supervisors/co -supervisors publications (RJSPc)	1.335	5.782	4	2

The following formulas were used to calculate the average of data in Table 4.2 :

$$TOp = Pp + Jp + JPc + JSPc \tag{4.1.2.1}$$

$$ROp = (\frac{Pp}{TOp}) * 100 \tag{4.1.2.2}$$

$$RJp = (\frac{Jp}{TOp}) * 100 \tag{4.1.2.3}$$

$$ROs = (\frac{Sp}{TOp}) * 100 \tag{4.1.2.4}$$

$$RJc = (\frac{Jc}{ToP}) * 100 \tag{4.1.2.5}$$

$$ROc = (\frac{Cp}{TOp}) * 100 \tag{4.1.2.6}$$

$$RJPc = (\frac{JPc}{TOp}) * 100 \tag{4.1.2.7}$$

$$RJSPc = (\frac{JSPc}{TOp}) * 100 \tag{4.1.2.8}$$

As noticed in Table 4.1, the mean of "supervisor publications" is higher than the mean of "co-supervisor publication" on Researchgate.net, owing to the limited number of Ph.D students (only fifteen with two supervisors). This is reflected in the percentage on the graph owing to the limited number of co-supervisor publications. These tables indicate two factors First, faculty supervisors collaborate more with graduate students and the resulting synergy translates into higher productivity. Second, there may be a reverse impact: namely, that graduate students, who need publications in order to graduate and begin their careers, select their faculty supervisors on the basis of their respective publication records. In other words, faculty members with more publications may well prove to be more attractive as supervisors. Euler and

Venn Diagrams were used, to represent joint publications between supervisor/Ph.D.student, co-supervisor/Ph.D. student and supervisor/co-supervisor. Figure 4.4 shows the relationship between them.

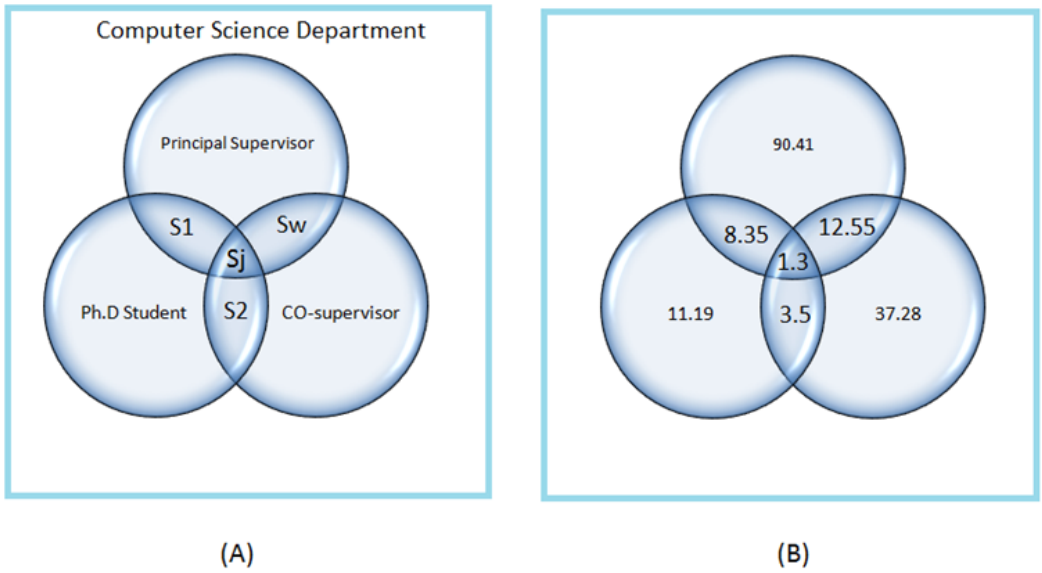


Figure 4.4: Euler and Venn Diagram for joint publications among Ph.D students, principal supervisors and co-supervisors.

Figure 4.4 (A), shows interconnected sets of circles where, we had a set of Ph.D students, a set of supervisors, co-supervisors which are considered part of a set of Computer Science Departments at Canadian universities on the Researchgate.net. (B) Shows joint areas (zones) among them, and the shaded zone represents the intersection among the three sets of circles, where there is collaboration among researchers in paper publishing.

4.2 Co-authoring among Canadian Researchers: Locally, Nationally and Internationally

Generally the trend of joint publication can be seen in Table 4.3. A list of Canadian computer science departments on Researchgate.net, was created to calculate the total number of publications. Researchers' profiles were read to find out the degree of collaboration among them. Joint authoring is to be investigated on local, national and international levels. Incomplete profiles were excluded and data was filtered, where only co-authored publications on a researcher's profile were considered. The final obtained data was categorized into four groups in an ascending order based on the number of publications.

Table 4.3: First group ranging from (1-800) of collaboration on "Local" , "National", "International" levels.

Univ.Dept.	Members	No.of filtered publications	Local%	National%	International%
Laurentian University	16	69	40.58	37.68	21.74
University of Northern British Columbia	13	92	31.5	33.7	34.8
Lakehead University	11	118	36.44	31.36	32.2
Brock University	12	118	26.27	34.75	38.98
Acadia University	10	123	18.7	23	58.54
The University of Winnipeg	18	146	23.97	35.62	40.41
University of Guelph	33	396	34.6	40.47	24.75
Ryerson University	47	794	33.74	18.2	48.06
University of Lethbridge	25	422	25.83	40.52	33.65
University of New Brunswick	39	523	30.8	30.01	39.2
Memorial University of Newfoundland	23	554	24.37	31.77	43.86
University of Regina	38	628	31.69	34.7	33.6
University of Windsor	56	632	34.5	28.8	36.71
Laval University	60	701	25.1	29.7	45.2

Table 4.4: Second group ranging from (801-1600) of collaboration on "Local" , "National", "International" levels.

Univ.Dept.	Members	No.of Filtered Publications	Local%	National%	International%
University of Manitoba	46	868	33.76	29.26	36.98
University of Western Ontario	57	908	33.15	23.99	40.86
Dalhousie University	78	1098	31.42	29.87	38.71
Universit du Qubec Montral	65	1201	24.064	36.719	39.217
University of Saskatchewan	71	1301	29.33	33.51	37.5

Table 4.5: Third group ranging from (1601-2400) of collaboration on "Local" , "National", "International" levels.

Univ.Dept.	Members	No.of Filtered Publications	Local%	National%	International%
Carleton University	65	1840	31.41	27.99	40.6
Universite de Montreal	91	2387	30.71	29.7	39.59

Table 4.6: Fourth group ranging from (2401-above) of collaboration on "Local" , "National", "International" levels.

Univ.Dept.	Members	No.of Filtered Publications	Local%	National%	International%
University of Victoria	81	2526	31.63	28.15	40.22
McGill University	105	2687	33.9	24.71	41.38
University of Waterloo	155	2733	29.75	28.8	41.46
University of Toronto	149	2875	29.67	24.94	45.39
University of Calgary	116	2993	29.946	36.409	33.344
University of British Colombia	119	3002	39.217	31.279	38.274

The above tables, shows the groups and the number of co-authored publications locally, nationally and internationally. Thirty one university departments were analyzed to find out researchers affiliation and co-authoring collaboration on different levels. Due to the departments varied contribution (universities with less than 10 members were discarded) such as the University of Prince Edward Island, which accommodates only (4) faculty members with (25) publications, compared to the University of Waterloo with (155) faculty members and (2733)

publications. Total filtered was (27 universities in a descending order) and tables are organized according to the total number of publications uploaded on researcher's profile on Researchgate.net. In reading the tables two points are to be considered. First, Researchgate.net is a scholarly collaboration platform with two aims, one is allowing self-archiving and the other is building scientific reputation. This is clearly seen in the case that some, well-established researchers list limited number of their publications or what they consider as their top publications. Second, the varied number of faculty presence on Researchgate.net out of the actual numbers of department faculty members, in addition to the dynamic nature of Researchgate.net. The tables were considered as an indicative reading.

It is clear that there is a high degree of collaboration between Canadian researchers and other researchers from around the world in the field of Computer Science. International co-authoring has the merits of allowing the research team a better chance of ideas exchange, especially across disciplines and expertise pooling enables researchers to handle complex and more visibility [38] and higher quality. Where they carried out a study on a dataset of 65 biomedical scientists at a New Zealand university. Collaboration variables for a 14 year period were coded and scientists detailed analysis revealed a positive relationship between the quality of articles and local (within-university) and international collaboration. The average percentage of within-university co-authoring among Canadian researchers was 29.74, nationally was 30.75 and internationally was 39.8. It can benefit researchers in developing countries by allowing them opportunities to interact with well established, recognized Canadian researchers. This will enhance their expertise and open new doors for publishing in peer - reviewed journals. It is fair to say that Canadian researchers benefit as well from other researchers and their available resources. The high percentage of international joint publications can be attributed among other reasons, to the new digital environment, the growth of academic social networks, the information flow on

the Internet, globalization of knowledge and collaboration.

4.3 Knowledge Sharing among Canadian Researchers on Researchgate.net

Collaboration takes different forms, and the importance of Questions/ Answers activities on Researchgate.net is essential. Answering questions gains special importance due to the fact that it is coming from a researcher's "tacit knowledge". This knowledge is rather "personal" including their expertise and skills which they have developed over the years. For published "explicit" knowledge, the known scholarly resources and Google take care of that. From this point answering research-related questions on Researchgate.net is of special importance. Data was collected by our crawler script, for a three months time-span to study Question/Answer activities, among researchers on Researchgate.net. One can say that asking a question is categorized as "Information Seeking" activity, while answering a question is a "Knowledge Sharing" activity, where Researchgate.net social network as a collaboration platform provides assistance. The total number of this activity is rather limited among Canadian researchers, reflecting limited use of Researchgate.net as platform for collaboration, and knowledge sharing in the field of Computer Science as can be seen from Table 4.7.

Table 4.7: Question/Answer on Researchgate.net.

Months	Mean of Questions	STDEV	Mean of Answers	STDEV
April	1.809	1.527	15.476	43.463
May	2.077	1.792	20.647	52.098
Jun	2.08	1.759	18.261	47.522

The total number of Canadian researchers posting questions on Researchgate.net is rather limited and Table 4.7, shows the average monthly score of a total number of (52) researchers,

after filtering out of the original data of (506). Answers come in two forms one as direct reply to a posted question and the other as a comment on certain questions in discussions. Answers represent the total number of answers by a researcher to questions posted on reserachgate.net, as shown on his profile. For questions in April (47) researchers posted questions with a total of (85) questions, followed by (52) in May with total of (108) questions and (50) in June with total of (104) questions. The total number of questions has decreased by 4 due to the cancelation of some of these questions, with few questions left unanswered (open). The top number of questions posted is (16) by a Ph.D student. For answers in April we had (42) researchers with a total of (650) answers, in May(34)researchers with a total of (702) answers, and in June (46) with a total of (840) answers. Some of these researchers are very active, for example, found was a researcher with more than (221) answers. This can be explained on the light of previous interpretations, that academic engagements, workload and probably, viewing Researchgate.net by some researchers as self-archiving platform more than a collaboration platform, stands behind that, in other words Canadian researchers are willing to participate in answering other researchers questions more than posting questions. They are more active in "knowledge sharing" than "information seeking", providing other researchers with useful ideas to enhance research activities.

4.3.1 Who Answers on Researchgate.net?

According to the questions data presented in Table 4.7, there is readiness for collaboration among the Canadian researchers. Data was collected to see the performance of Canadian researchers in terms of collaboration among them and globally on the academic social network Researchgate.net in terms of question/answer. The average number of answers to a researcher questions was (16.577) with Stdev (44.421). The average of the top five countries from which

answers came are:

1. India with average of 4.423
2. Canada with average of 3.385
3. USA with average of 2.423
4. Germany with average of 1.692
5. France with average of 1.038

The average of answers coming from Canadian researchers in Table 4.8 compared to answers coming from different researchers globally.

Table 4.8: Answers from Canadian researchers and different countries

Country	Mean of answers
India	4.423
Canada	3.385
U.S.A.	2.423
Germany	1.692
France	1.038
U.K.	1
Pakistan	0.577
Spain	0.423
Turkey	0.346
Brazil	0.269
Finland	0.269
Portugal	0.231
Italy	0.192
Mexico	0.192
Israel	0.154
Malaysia	0.154
Japan	0.154
Chile	0.115
Australia	0.115
Switzerland	0.115

Country	Mean of answers
China	0.115
Netherlands	0.115
Croatia	0.115
Sri Lanka	0.077
New Zealand	0.077
Philippines	0.077
Denmark	0.077
Sweden	0.077
Jordan	0.077
Greece	0.077
Poland	0.038
Egypt	0.038
Colombia	0.038
Serbia	0.038
Bulgaria	0.038
Georgia	0.038
Saudi Arabia	0.038
Russia	0.038
Iran	0.038
Thailand	0.038
Czech Republic	0.038

In Table 4.8, the Canadian researchers ranked second after the Indian researchers in answering questions on Researchgate.net. This indicates the willingness of Canadian researchers to collaborate and to share their knowledge. Detailed analysis revealed that:

1. Discussions related answers to posted questions were taken in consideration through the analysis.

2. Top Canadian researchers participating in posting questions or answering ones came from the University of Manitoba.

3. Independent researchers, who have no affiliation to any institute or university were counted according to their countries.

4. Post doctoral fellows came at the top of Question/Answer activity on Researchgate.net

and this finding is consistent with Almousa [39] findings by his study for three months in 2011 on members from five disciplines including Computer Science on Academia.edu. Collected data was codified and analyzed showing that the top group in asking and answering question activity was Post Doc. group 0.008 compared to other faculty 0.006. He noticed that this group activity is lower than the same group activities of other disciplines.

5. Some questions were left unanswered for a long time marked as (open). Ph.D students and post doctoral fellows, got multiple answers to their questions, for example a single researcher answered (221) to questions posted by other researchers.

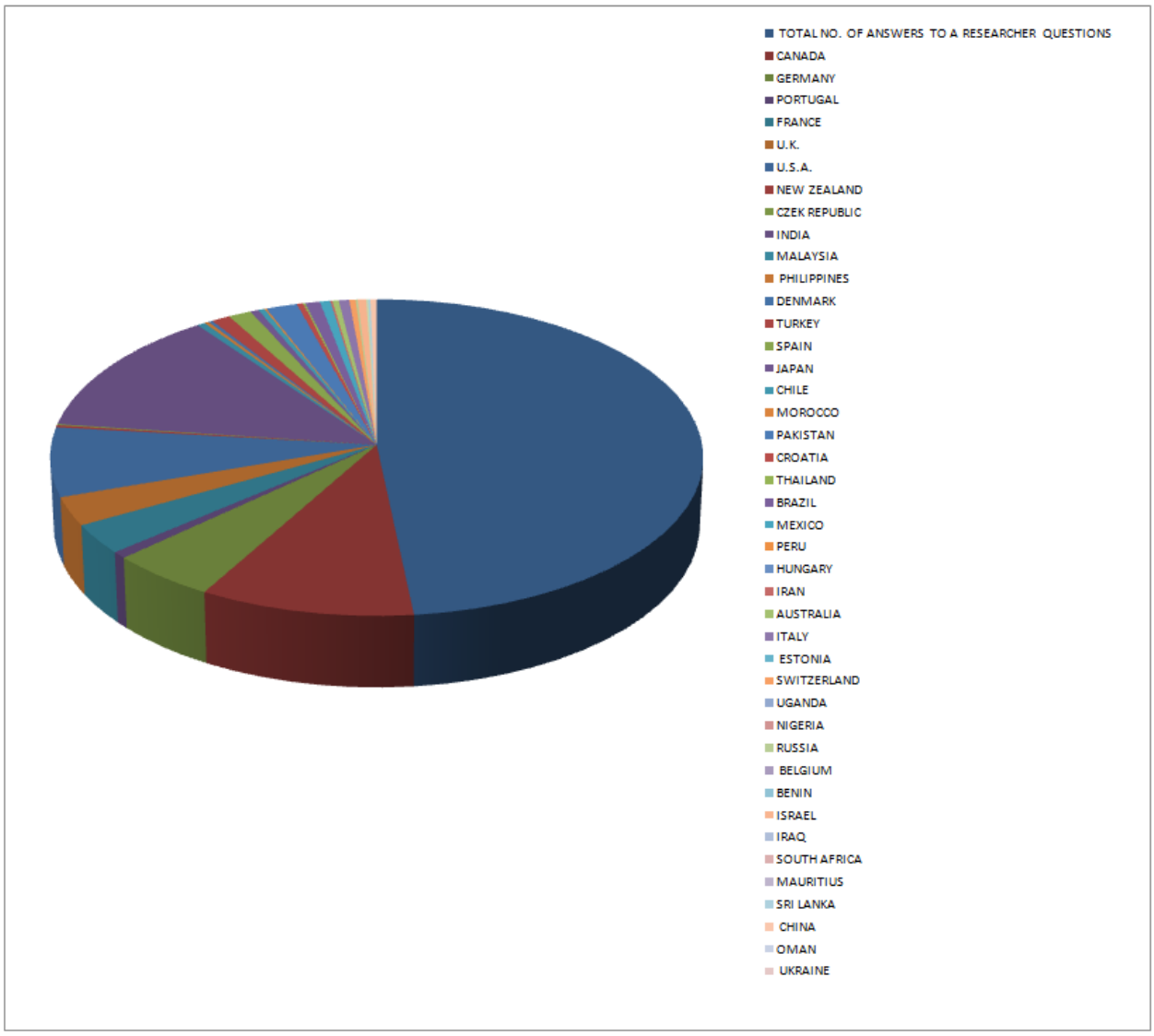


Figure 4.5: Countries ranking in terms of Question/Answer activity.

It is notable that co-authoring is not the only aspect of collaboration offered by Researchgate.net, but it allows for self-archiving, reputation building and more visibility. Materials in different formats, data-sets, articles, proceedings, technical reports, patents, chapters, books, theses, and even negative results are uploaded by a researcher, contributing to "Open Access" movement. Activities were investigated of the researcher engaged into Question/Answer ac-

tivities to identify his/her other activities. Table 4.9 shows that but, very limited number of patents was found and consequently it was disregarded.

Table 4.9: Performance of different formats materials upload on Researchgate.net.

	Mean	STDEV
Articles	7.826	9.896
Chapters	1.75	1.035
Conference Papers	8.842	11.377
Datasets	1.8	1.704
Full texts	12.214	14.247
Books	1.75	1.389

The population of (52) researchers were not only active in questions and answers activities, but they were active as well in contributing materials to the research community, such as (Articles, Chapters, Conference Papers, Datasets, Full Texts, Books). The top total number of published articles within the sample was (34) by Adjunct Professor, followed by (26) articles for a professor, with limited activities by graduate students. For chapters, the highest number was (3) by the same adjunct professor. Regarding conference papers the highest was (39) posted by the same adjunct professor, while Ph.D students contributed more datasets on the Researchgate.net with the highest number of (6). Professors contributed top contribution was (45) Full Texts, and (4) books with limited number of theses (8), and only (2) patents were shared. Since a very limited number of other formats was observed they were excluded. Table 4.9, indicates that only limited number of Canadian researchers in the field of Computer Science are engaged in knowledge sharing activity (52) compared to the total number of the main sample of (506). Besides the selected sample of (52) showed activities in uploading different materials formats.

4.4 Researchgate.net Metrics

Further analysis was done from the metrics perspective, Researchgate.net came with different metrics “Impact Points” and ”RG Score. Metrics were examined on Researchgate.net to find out if there is a correlation between Followers/Views, Followers/Downloads, Publications/Views and Researchers rank /Downloads. Two different software were used, Excel to find the value of(r) and Minitab17 to find ANOVA table with Regression Analysis for more details. The relationship was investigated between:

- The number of Followers and Downloads.
- The number of Followers and Views.
- The number of Publications and Views.
- The number of Downloads and Researcher’s Rank.

4.4.1 Correlation between Followers and Downloads

To find the relationship (Correlation) between Followers and Downloads, a linear Regression Model was used. Before discussing the correlation between these two variables, this model had to be defined as to what to do exactly. A Linear equation can be used for data observation to model the relationship between two variables, one is the explanatory variable and the other is the dependent one. It is necessary to determine the presence of relationship between the variables prior to fitting a model for data observation with the existence of a noticeable association between the variables. In case of the absence of any association between variables, then using a linear regression model to the data might not be of value. Correlation and coefficient (between -1 and 1) a numerical measure of association is useful for measuring association on the line of linear regression. The equation in the form of $Y = a + bX$, is used (X represents the Explanatory

variable, and Y represents the dependent one). b is the line slope and a is intercept (Y value when X=0).

In our case we have Y as a number of Downloads and X is the number of Followers.

Y=Downloads.

X=Followers.

The following formula of (r) correlation coefficient is:

$$r = \frac{1}{1-n}\sum_{n-1}^{506}(\frac{X_i - \bar{X}}{S_x})(\frac{Y_i - \bar{Y}}{S_y}) \tag{4.4.1.1}$$

$$b = r\frac{S_y}{S_x} \tag{4.4.1.2}$$

$$a = \overline{Y} - b\overline{X} \tag{4.4.1.3}$$

If the value of r is positive, this means there is a positive relationship between the number of Followers and the number of Downloads (increased value of one variable corresponds to increase in another variable). If the value is negative, there is a negative relationship between the two variables (increased value of one corresponds to decrease of another one). In addition, we used ANOVA table for the variance analysis.

Using ANOVA Table

ANOVA Table describes the complete analysis of variance between two variables as we are finding here, the relationship between the variables. Where Sum of Squares (SS), and Mean of square (MS) are required to find the value of F Ratio. By using the value of F Ratio, we can conclude about the variation of two variables used to find the relationship. Errors tells us how

fit the Scatter Data with Fitted Line. If the value of Error is greater, then the relationship is minor and if the value of Error is Less, then there is strong relationship. The equation of r has been calculated by using MS. Office Excel, to compare the value of r with the one obtained from the third party Statistical Tool (Minitab17). Table 4.10, shows the first calculation:

Table 4.10: Finding r between "Followers" and "Downloads".

$\overline{X}$	$\overline{Y}$	S_x	S_y	r
17.334	428.194	23.397	1393.092	0.503

Regression Analysis was made using Minitab17 and the regression equation is:

Downloads = - 92.20 + 30.02 Followers

R-Sq = 25.4% R-Sq(adj) = 25.3%

The value of R-Sq will always pass through the mean of X and Y. The regression line has to be described by the mean, standard deviations, and correlation of two variables.

Table 4.11: Analysis of Variance "Followers" and "Downloads".

Source	DF	SS	MS	F	P
Regression	1	249152285	249152285	171.80	0.000
Error	504	730904246	1450207		
Total	505	980056531			

The degrees of freedom as in Table 4.11, denoted here as (DF), which is a number of ways, to be given to our sample freely. The total data points collected (n), and to calculate total degrees of freedom n-1= 506-1 =505. If there was a group of data that need to be compared (m) in this case the Degrees of Freedom associated with the regression is (m-1)= 2-1=1, according to the

two columns of data "Followers" and "Downloads". Since we had (n) of data points collected with (m) of groups of data, which needs to be compared, the Error Degrees of Freedom is n-m= 506 -2 =504. To calculate the Sum of Squares (SS), we had to know that total variation in the data consisting of two components, one related to regression, and the other related to random error. To calculate SS (Total) the following formulas were used [40]:

$$\bar{X_i}. = \frac{1}{n_i} \sum_{j=1}^{n_i} X_i j \tag{4.4.1.4}$$

Mean of the data for group i (Followers & Downloads), where i = 1,2

$$\bar{X}.. = \frac{1}{n} \sum^{m} \sum^{n_i} X_i j \tag{4.4.1.5}$$

Mean of mean of (Followers & Downloads)

$$SS(Total) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_i j^2 - n\bar{X}..^2 \tag{4.4.1.6}$$

Total sum of squares used to measure the differences in the data without regard to its source.

$$SS(B) = \sum_{i=1}^{m} n_i \bar{X_i}.^2 - n\bar{X}..^2 \tag{4.4.1.7}$$

SS(B) to measure the differences between the effectiveness groups

$$SS(E) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_i j - \bar{X_i}.)^2 \tag{4.4.1.8}$$

The error sum of squares were used to measure the differences in the data, which is the sum of

squared distances of X_ij to the means $\bar{X_i}$.

$$MSB = \frac{SS(B)}{m-1} \tag{4.4.1.9}$$

while

$$MSE = \frac{SS(E)}{n-m} \tag{4.4.1.10}$$

$$F = \frac{MSB}{MSE} \tag{4.4.1.11}$$

The F ratio is the ratio of two mean square values, if F is large it means that the difference among group means is high, and in our case it is “Followers” and “Downloads”. The data sample collected as in Table 4.11, was random data ended up with large values in some groups and small in others.

P - value computed from F ratio of ANOVA Table, and the two values of Degrees of Freedom. If

$$P - value < 0.01 \tag{4.4.1.12}$$

, there will be strong evidence against the hypothesis that says the difference in the means is due to the randomly selected data, as it was noticed from the scenario of “Followers” and “Downloads”. It is not necessary that all means are different from each other, but only one different from the rest is enough. The value of r =0.50 as in Figure 4.6, shows the cluster of dots are not approaching the straight line, therefore there is a moderate correlation which is considered a positive relationship between “Followers” and “Downloads”. The more followers a researcher has on Researchgate.net, the higher his downloads are.

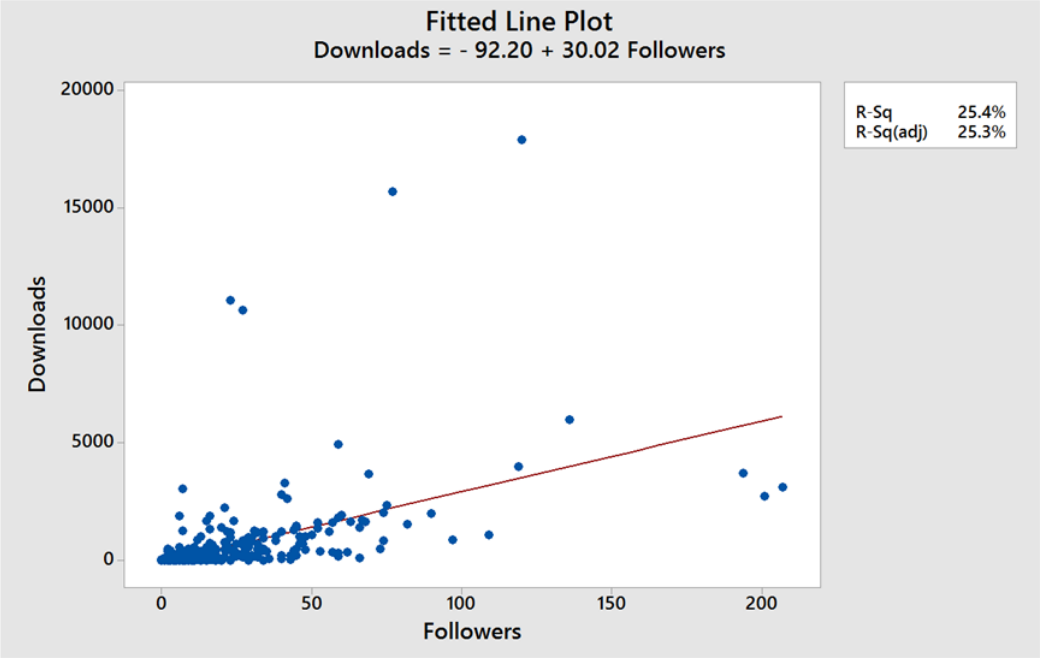


Figure 4.6: "Followers" versus "Downloads".

4.4.2 Correlation between Followers and Views

To find the relationship between "Followers" and "Views", data was organized on Excel sheet and before applying the correlation as noticed on Researchgate.net, "Views" of a researcher have the letter (k) ex.(3k) instead of figures, which interrupts our analysis. Therefore this letter was converted into(000). Table 4.12 shows the calculation of correlation coefficient where, X is for "Views", and Y is for "Followers"

Table 4.12: Finding r between "Followers" and "Views".

$\overline{X}$	$\overline{Y}$	S_x	S_y	r
1173.134	17.334	1910.087	23.397	0.665

Regression Analysis was made for more details and the regression equation is:

Followers = 9.694 + 0.000770 Views

$R - Sq = 44.6\%$ $R - Sq(adj) = 44.5\%$

Table 4.13: Analysis of Variance "Followers" and "Views".

Source	DF	SS	MS	F	P
Regression	1	123185	123185	405.12	0.000
Error	504	153253	304		
Total	505	276439			

The analysis of variance in Table 4.13: indicates that DF=1, which means that there is only one way for giving our sample freely. SS= 123185 and MS= 123185 have been calculated with reference to (Using ANOVA Table). F ratio: shows a large value in Table 4.13, and this indicates data random sample for "Followers" and "Views". The means of groups were varying due to the large values in some groups and small in others. P - value: Table 4.13 indicates small value of P, since one of the mean should be different of the rest, and even if the means are equal or different from each other. Complete evidence is not support that the random sampling is the main reason. The dots in Figure 4.7, are slightly far away from the straight line and the value of r= 0.668 showing the presence of moderate positive relationship, between "Followers" and "Views".

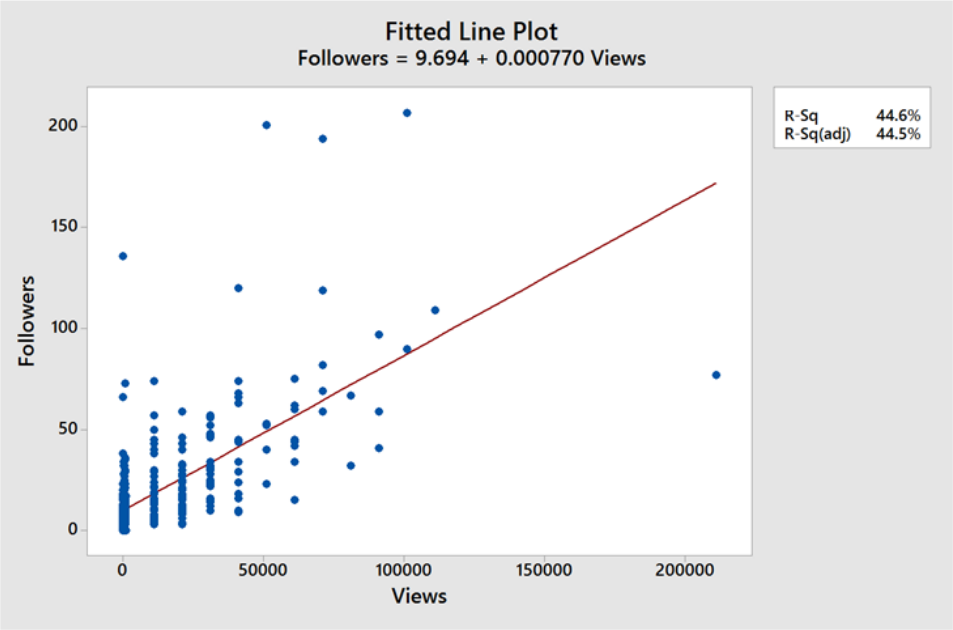


Figure 4.7: "Followers" versus "Views".

4.4.3 Correlation between Publications and Views

Usually a researcher has visitors to his profile, since this profile services purposes such as archiving the researchers publications, reputation building and visibility. A visitor has some queries in his mind for example what degree of expertise does a researcher have, what are his/her publications and contributions, and what is his/her standing among other researcher on Researchgate.net. Visits might result in following a researcher or viewing his/her publications or contacting this person. Here we studied whether a relationship exists between visits to his/her "Views" and the "total number of his/her publications".

Table 4.14 shows the calculation of r in MS Excel, where X is for Publications and Y is for Views

Table 4.14: Correlation between "Publications" and "Views".

$\overline{X}$	$\overline{Y}$	S_x	S_y	r
26.676	1173.134	45.083	1910.087	0.893

Regression equation is:

Views = - 774.1 + 400.7 Publications

R-Sq = 79.4% R-Sq(adj) = 79.4%

Table 4.15: Analysis of Variance "Publications" and "Views".

Source	DF	SS	MS	F	P
Regression	1	1.648	1.648	1945.49	0
Error	504	4.2703	8.473		
Total	505	2.075			

Table 4.15: indicates that F ratio is still high in this scenario due to the differences in the means and the random sample collected of "Publications" and "Views". P is small here due to incomplete evidence of these differences.

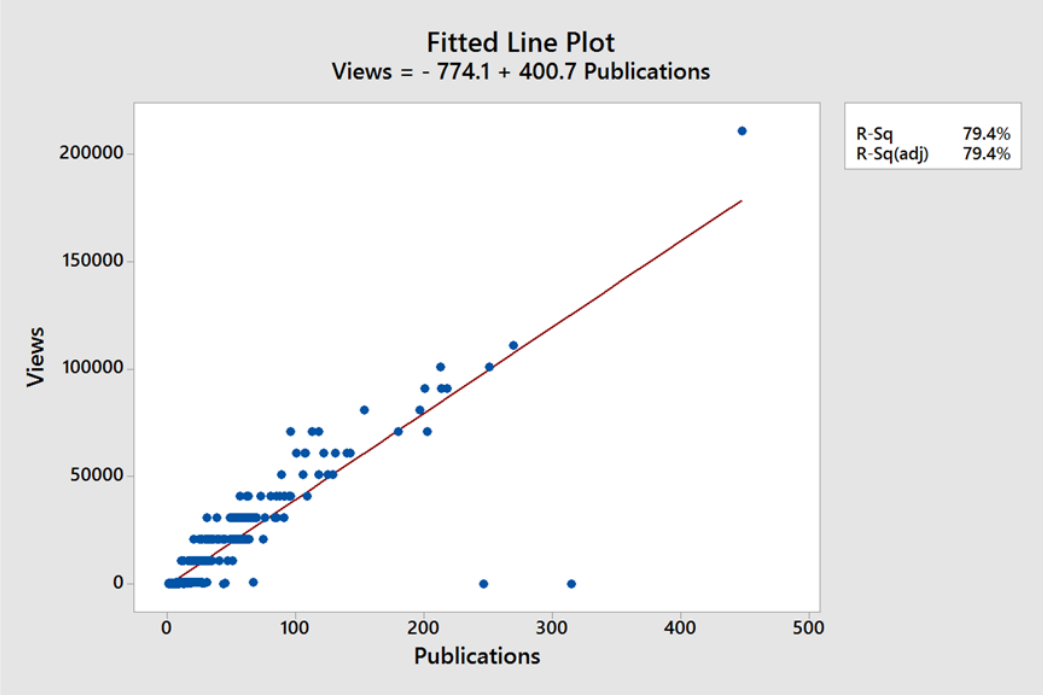


Figure 4.8: "Views" versus "Publications".

The dots in Figure 4.8, are close to the straight line. The value of r = 0.89, pointing to a highly significant strong positive relationship between the number of publications of each researcher and the number of Views.

4.4.4 Correlation between "Views" and "Number of Authors Per Paper"

The general views to the profiles have been showed previously in different scenarios, regarding the total number of Views vs Number of Publication and the number of Followers vs Number of Views. In this scenario data collected in details including Downloads, Views, Date of Publishing,

Authors on their byline positions from a sample of (458) publications and filtered to find the number of authors per publication.

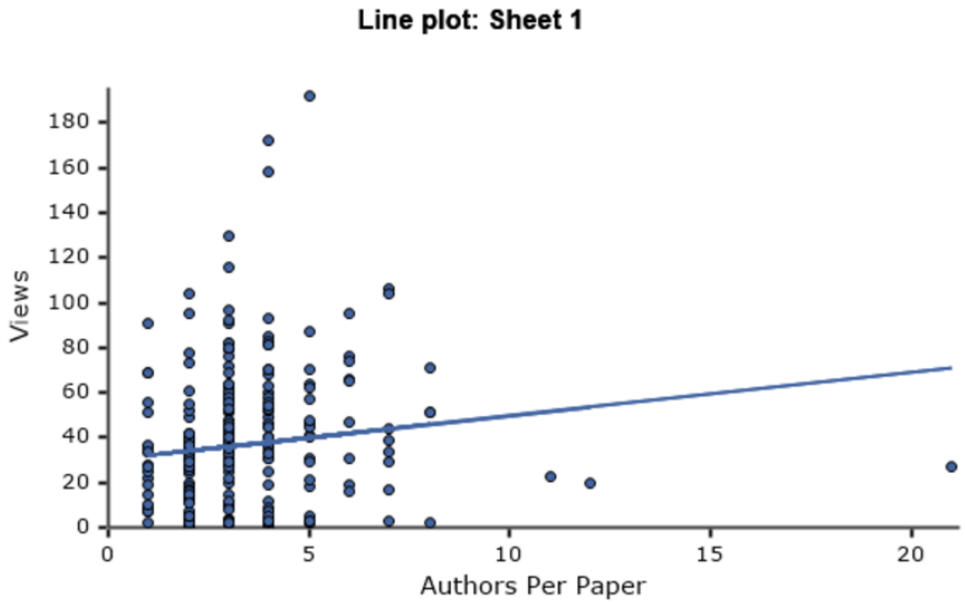


Figure 4.9: "Views" versus "Authors Per Paper".

The relationship between Views of any publication is not affected by the number of authors and the correlation between them is very weak with r= 0.125 as in Figure 4.9. The reader mostly relies on the contents of the publication rather than the number of contributors.

4.4.5 Members of Department versus Authors of Publications

Our work was developed to include a study (At a level of Departments) of different Canadian Computer Science Universities to (32) departments of our list of study, in addition to the previous study of individual researchers. This study focuses on the interaction of researchers on the department level, and departments data was collected and filtered. During filtration, it was noticed that not all the members of the computer science on researchgate.net are authors with posted publications, some have limited members such as university of Prince_Edward_Island.

In the following Figure 4.10, the correlation between the number of members and the authors of departments' publications was studied.

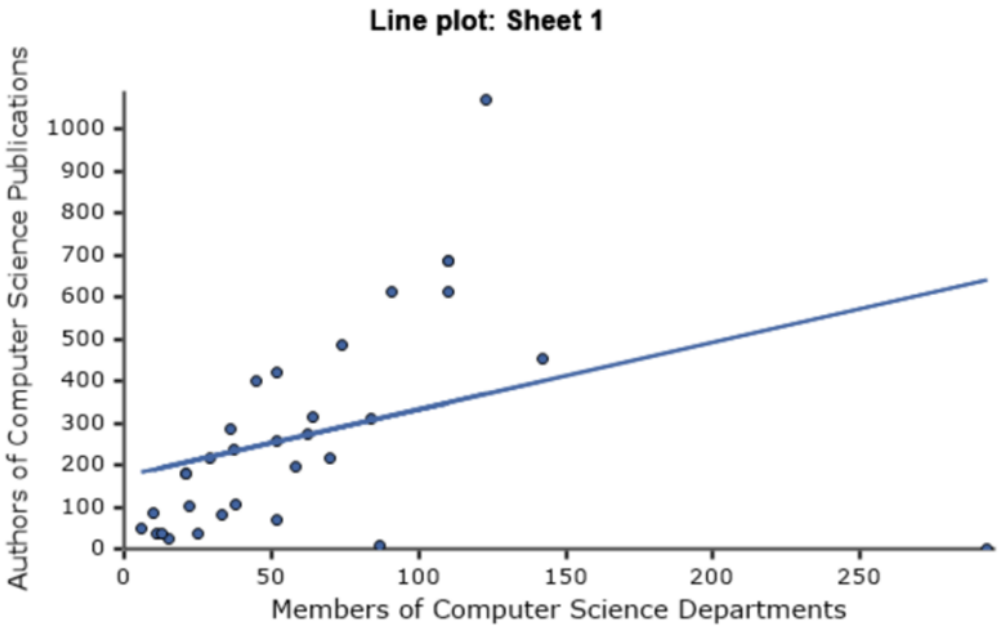


Figure 4.10: "The Number of Department's Members" versus "The Number of Authors of Publications of Computer Science Department".

The correlation r= 0.355 indicates almost a moderate relationship between the two. Some studies consider (r) less than 0.5 as a week relationship, so there is no significant indication that if the number of members increased, it would lead to an increase in the number of authors per department.

4.4.6 Correlation between Departments Publications and the Total Number of Impact Points

The calculation of impact points of department is based on total researchers' impact points posted on their walls, and the average impact points of department is the (Total of publica-

tions/Total impact points) The following Figure 4.11, shows the correlation between publications and total impact points.

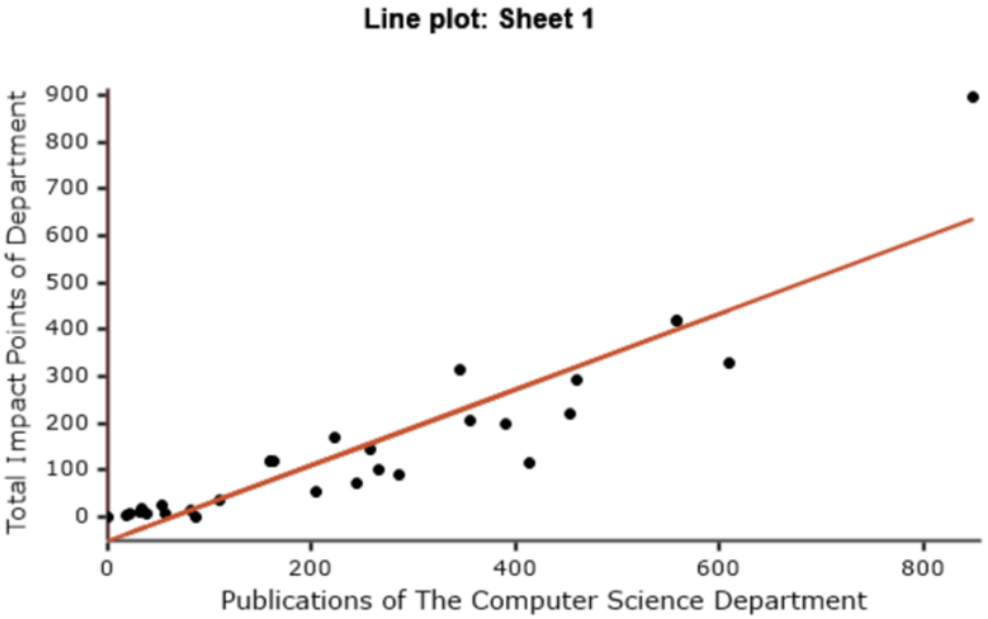


Figure 4.11: "Publications" versus "Total Impact Points".

r= 0.913, indicates a strong relation between the two. Researchgate.net takes time to update impact points and RG score values, if a researcher added a new publication, Researchgate community would see the journal in which his/her published with Thomson Reuters. Data was collected about the Computer Science departments of the (32) universities for a period of three months start from April-June. The development of the department's altmetrics displayed in Figure 4.12. Detailed analysis reveals that the top number of Computer Science members were from University of Waterloo with (147) members, while University of Toronto was ranked at the top in the number of publications with (849), total impact points with (896.82). Top Avg. impact points was for Universite de Montreal with (0.91) Avg. Highest number of publications authors came from university of Toronto with (1.071) authors. Which is a positive indicator of the of Computer Science members interactivity, and their willingness to join a research social

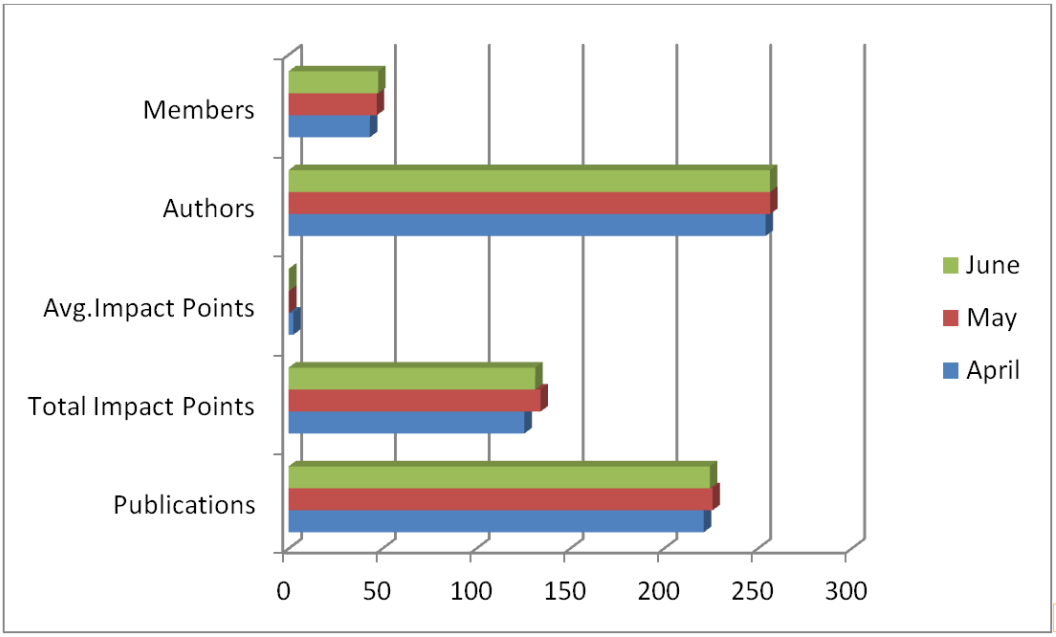


Figure 4.12: Developments of Departments altmetrics.

network for scientific purposes.

4.5 Altmetrics Influence

To study the altmetrics' influence on the data categories, different researchers categories (Faculty members, Postdoc. fellow, Ph.D students and MSc. students) were identified. The average, and Std Error of the four groups were found. According to the massive jump in the number of downloads specially for the faculty members, data varied between less than 10 to higher than 17000 downloads for a researcher. In this case the high numbers were dropped off the focus was on the cluster of data based on their mean in order to get a significant result. The main objective is to see if a researcher's rank has any influence on the number of downloads, or whether the reader relies on the contents of a publication regardless of the author's academic rank. Figure 4.13, shows the results.

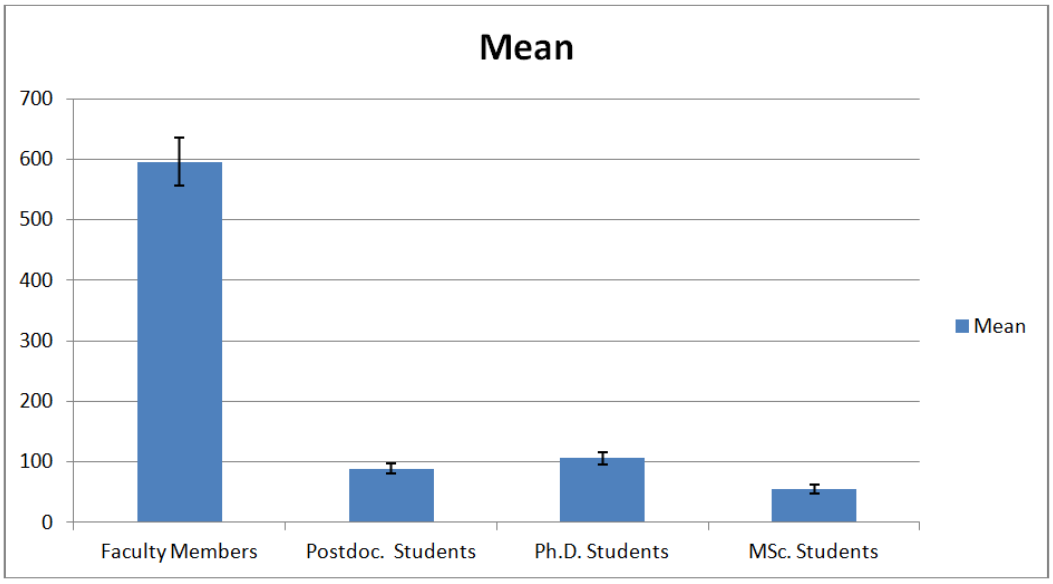


Figure 4.13: Box Plot for "Downloads" and "Researchers' Ranking".

The Std Error has been calculated based on the following formula: $STDErorr = \frac{Std}{SQRT(no.ofrows)}$. Despite the fact that Postdoc. are having a higher rank than Ph.D students, their downloads were less, but it was noticed that faculty members with high downloads was tied to their academic ranking. It is clear that using rank vs downloads yardstick, it does matter for faculty members but it does not for (Postdoc.,Ph.D,MSc.) students. It can be said that researchers on Researchgate.net download publications depending on the authors reputations and the quality of publications as well. The four categories with their publications are represented in Figure 4.14. The study showed that despite of post doc. students with higher number of publications than Ph.D students, but their downloads are less according to the previous study.

If the number of citations of these four categories is compared, the result would be clear as in Figure 4.15, that citing a paper can be affected by researcher's academic ranking. In the line the "Followers" for each category were tested as in Figure 4.16. This can be explained by the increasing numbers of publications for faculty members and consequently the rest of categories had increasing number of citations and followers.

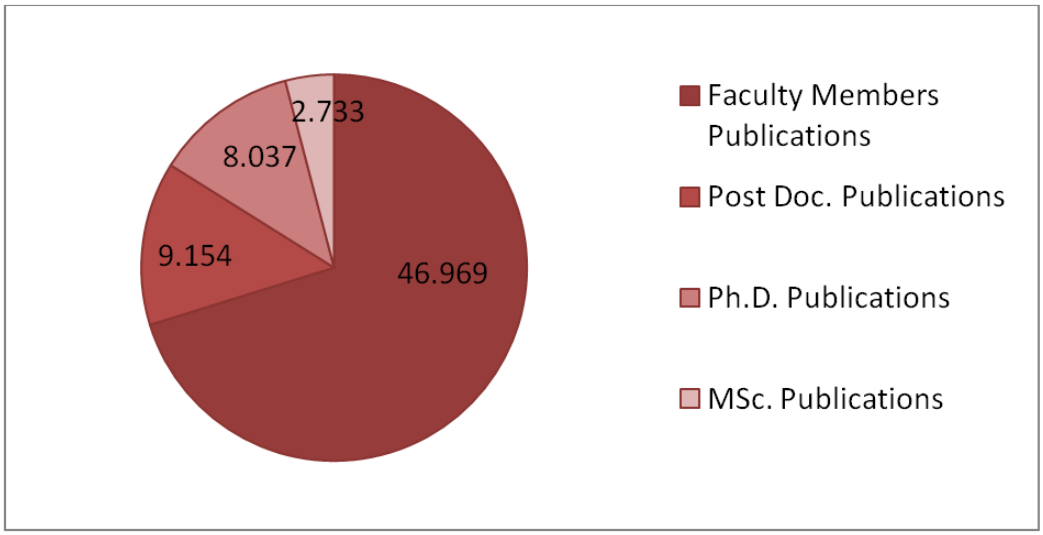


Figure 4.14: Data Categorized based on The Number of Publications.

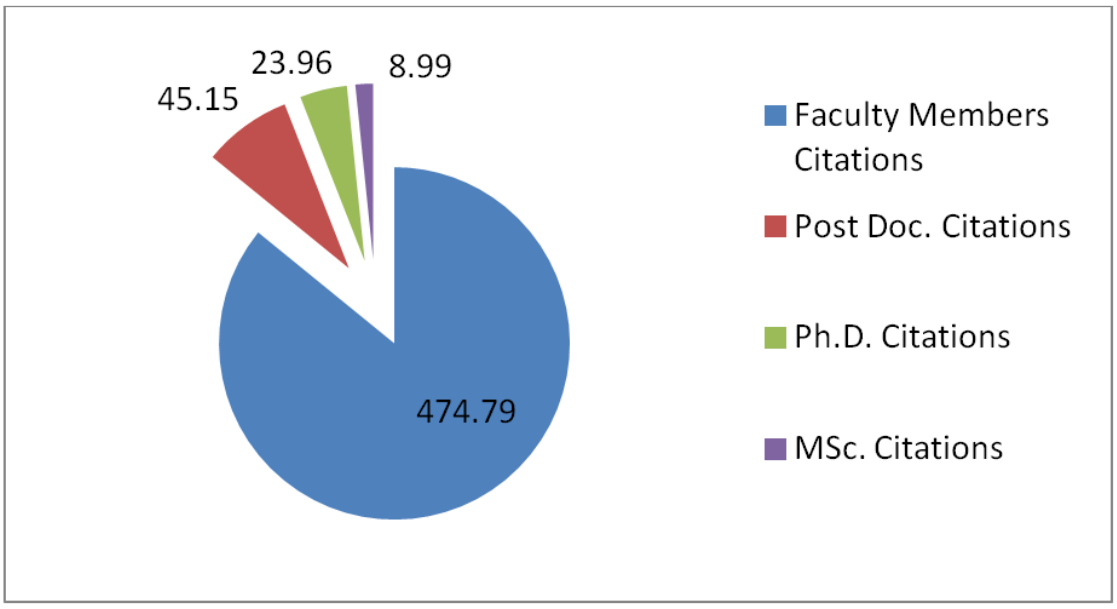


Figure 4.15: Data Categorized based on The Number of Citations.

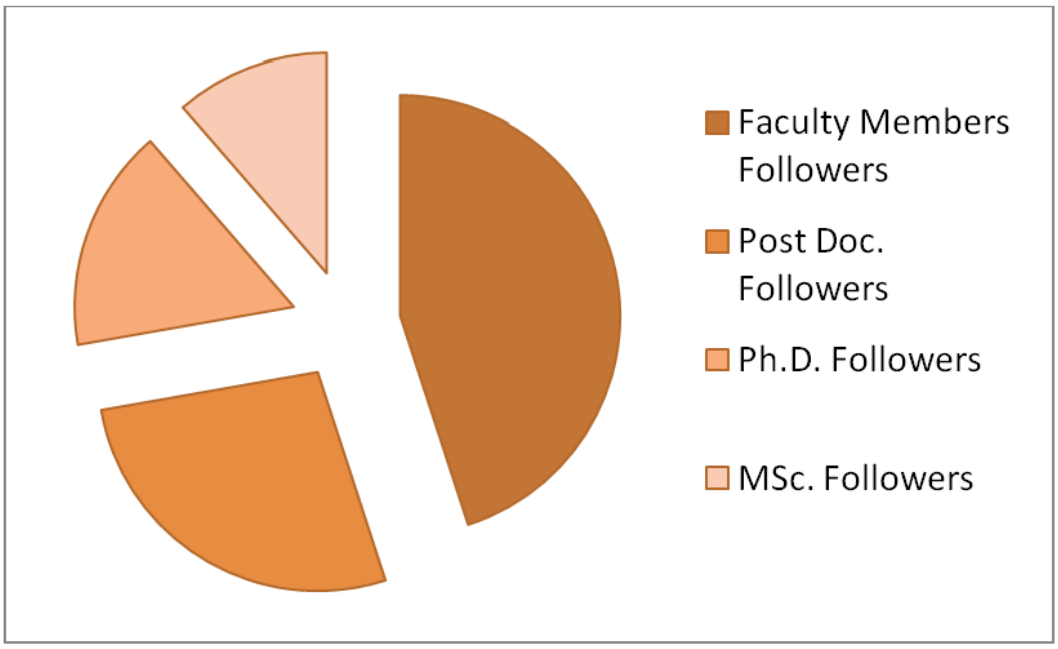


Figure 4.16: Data Categorized based on The Number of Followers.

4.6 Metrics Changes

A bi-weekly run of the crawler was made for three months starting on 3rd.April to 28th.June 2014. Since it was dealing with a dynamic academic social network, a researchers presence on Researchgate.net is not regular and statistics change.

4.6.1 "Followers" and "Publications" on Reserachgate.net

Researchers' publications update was monitored, and calculations were done on monthly bases. Table 4.16 shows the mean and standard deviation of "Followers" and "Publications" for April-May-June 2014.

Table 4.16: "Followers" and "Publications" readings.

	Mean of Followers	STDEV	Mean of Publications	STDEV
April	16.78	25.019	25.464	43.3659
May	17.332	25.709	25.546	43.384
June	16.71486	24.788	25.242	43.144

As noticed from Table 4.16, the average number of Followers went up in April - May, but went down consistently in June. This can be attributed to the fact that a researcher has full control on his profile and is free to upload or delete publications any time, and even some
completely removed their accounts or initiated new ones. A sample of (502) was taken due to the removal of four accounts details by the researchers.

4.6.2 "Views", "Citations" and "Downloads" on Researchgate.net

The growth of "Views", "Downloads" and "Citations" was monitored for the same period as shown in Table 4.17

In Table 4.17, there is a fluctuation of "Views" , "Downloads" and "Citations" from month to moth up and down. This reflects different activities of researchers from time to time probably this can be attributed to different academic responsibilities, workload or other reasons.

4.6.3 "Impact Points" and "RG Score" Change on Researchgate

There are a growth of the "Impact points" and the "RG score" of researchers on Researchgate.net. "RG score" calculation is done based on four different elements "Publications", " Questions", " Answers" and "Followers". Researchgate.net community contributes to the points of a researcher, based on how they received his publications, and his score is incremented accordingly. Researchers interaction with individual publication whether through downloading, requesting, viewing or commenting his publications are considered. Calculating impact points by Researchgate.net, takes in consideration the journal impact factor as presented by (ISI), and

we monitored changes occurring in the "impact points" scoring. The impact points indicate the total impact factors for all the journals, so if the researcher has published two papers in journal A and three papers in journal B, his total impact points will be A+A+B+B+B. Table 4.18, Shows changes in impact points and RG score on Researchgate.net.

Table 4.18: "Impact Points" and "RG Score" on Researchgate.net.

Month	Mean of Impact Points IP	STDEV	Mean of RG Score	STDEV
April	6.347	14.994	3.576	6.045
May	11.868	116.207	4.118	6.254
June	11.91	116.095	2.959	5.602

As noticed from Table 4.18, that more high impact factors publications are uploaded to the researcher profile and more interaction with his/her publications were made. Also it points out that other Researchers interaction went up, before plunging down in June, and probably that's due to possible time pressure and carrying out other academic activities. It can be attributed to the fact that the majority of researchers are concerned with the self- archiving aspect of Researchgate.net. This might result in less frequent visits and less interaction on Researchgate.net.

4.7 Author's Position

Background Information

The traditional method was to look at a researcher's achievements through his publications in peer-reviewed scholarly journals. The Journal Impact factor is taking into consideration when it comes to promotion, tenureship or funding. Researchgate.net came with new altmetric tools for evaluating researchers performance based on (ISI) journal factor, and the researchs activities in addition to how the scientific community received his/her publications. It is thought that by adding a new dimension to that altmetric it will be useful in giving a more meaningful

reading. As known in calculating the journal impact factor (ISI) it does not consider the authors sequence in its calculations and it is suggested a method for authors sequence determination and to be taken in calculating the impact factor. Keeping things simple, is the guiding principle in this scheme design. It is fully realized that discussing and handling the Impact Factor is beyond the scope of this thesis and it is part of another field. A crawler script was written as mentioned in algorithm (10) in Appendix A, to collect detailed data about the publications of the researchers. Data was organized on Excel sheet under the titles ("University Name", "Departments", "Title of Publication", "date of publishing", "Main Author", "First Author", "Second Author",....., "Twenty Author"). The challenge, faced is that some researchers have more than one profile for unknown reason. Calculations were made after profiles verifications, and in case of duplication for the same profile, only one was considered. In order to study the authorship positions, we had to know the number of times each researcher came as main author, first co-author, second co-author, third co-author, ... , sixteen co-author. The highest position number found on Researchgate.net among Canadian Computer Science researchers, was twenty. As mentioned in algorithm (11), Appendix A, an array has been created to read 20 positions with a counter to start reading these positions. The crawler retrieved these names in columns, and it was encoded into the numbers using some mathematical functions on Excel. The major contribution in multi authored papers is assigned to the main author, first, second and third co-authors, according to the Sequence Determines Credit (SDC) model, which was accepted by us since it was greatly accepted informally by Canadian researchers. Correlation between the number of publications and the authors byline positions is going to be calculated accordingly to find the relationship between them. Figure 4.17, shows the correlation between the total number of "Publications" and "Main Author". Linear regression model, has been used and the regression equation is:

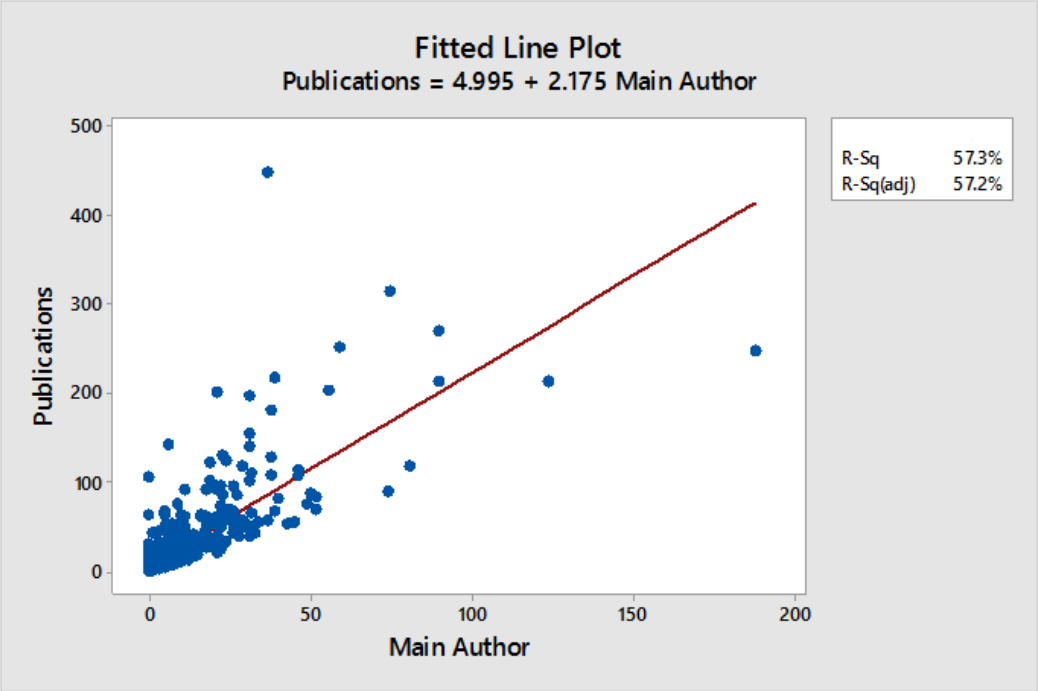


Figure 4.17: Correlation between Publications and Main author position.

Publications = 4.995 + 2.175 Main author

The coefficient of correlation r = 0.756 indicating a significant relationship between publications and main author position. Other studies showed increase in the Canadian academics publications after 50 years age until their retirement. The referred as well, that they move closer to the end of byline as they grow older, but their scientific impact increases Gingras etal..... [41].

4.7.1 The Relationship between Total of Publications and 1st Co- Author

The same approach is applied to study the number of publications and the opportunity of falling into the 1st. co- author position.

The Linear regression model is:

Publications = 5.130 + 2.507 1st.Co-author

R-Sq = 85.1% R-Sq(adj) = 85.1%

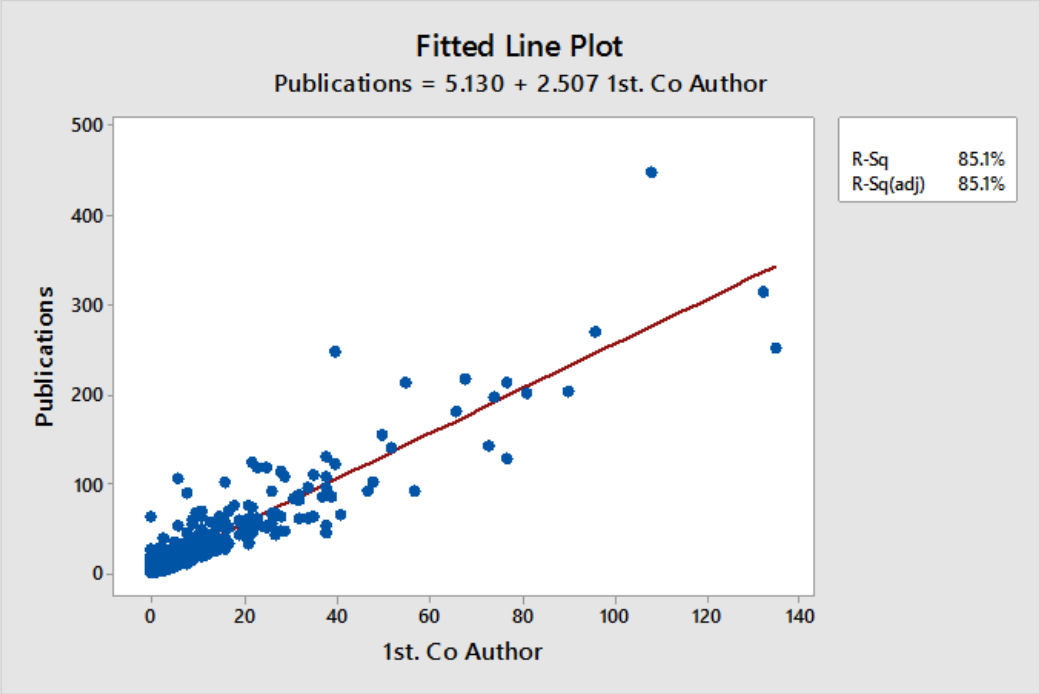


Figure 4.18: Correlation between Publications and 1st. Co-author.

The correlation coefficient r = 0.922, shows that there is a very strong relationship between the total of publications and 1st. co- author position. This study findings are consistent with previous study conducting by Gingras, et al [41] on a sample of 6,388 university professors in Quebec in 2008 indicated that older professors tend to move a way from main author in their publications.q

4.7.2 The Relationship between Total of Publications and 2nd. Co- Author

The regression equation is:

Publications = 12.26 + 3.048 2nd. Co-author

R-Sq = 72.4% R-Sq(adj) = 72.3%

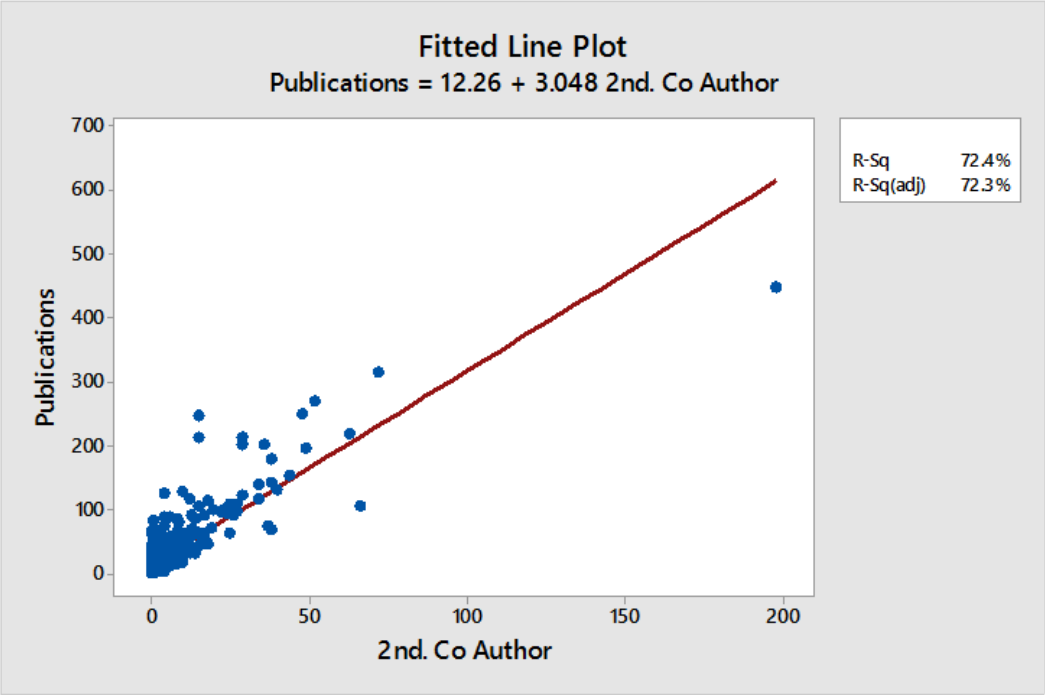


Figure 4.19: Correlation between total of Publications and 2nd Co-author.

The correlation coefficient r = 0.850, shows strong a relationship between publications and 2nd. co-author and the data is not fitted as compared to 1st co- author. Through browsing into the collected data, it was noticed that the presence of authors with more than 500 (one or two), publications falling late on the byline position, in which they have been disregarded. The highest number considered in our analysis was (448) publications.

4.7.3 The Relationship between Total of Publications with 3rd Co-author

The regression equation is:

Publications = 14.63 + 6.394 3rd.Co-author

R-Sq = 65.3% R-Sq(adj) = 65.2%

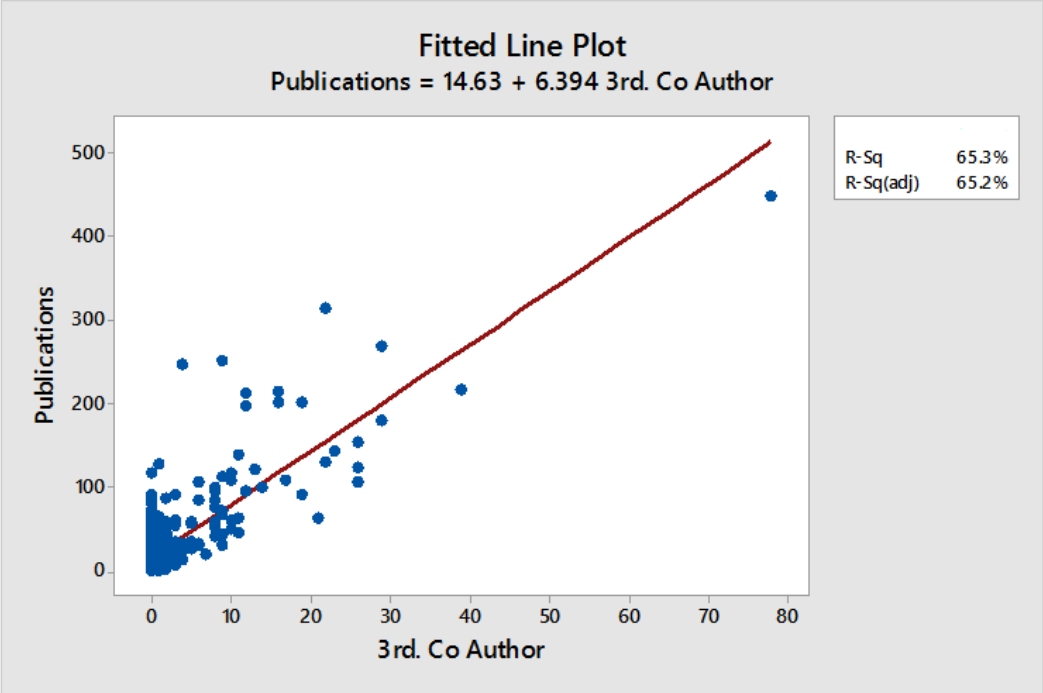


Figure 4.20: Correlation between total of Publications and 3rd.Co-author.

The coefficient of correlation r = 0.808 here, shows a strong relationship between publications and 3rd co- author position. In general the Canadian authors keep being main contributors as they grow older and fall within the first quarter of multi-authored papers, where the phenomenon of the "Free Rider" or "Guest Author" noticed in other disciplines does not exist.

4.8 Summary

This chapter discussed the degree of co-authoring between supervisors and Postgraduate students on the levels of Ph.D. and MSc. is revealed. Co-authoring among Canadian researchers in the field of Computer Science is investigated on local, national and international levels. The correlations between different metrics was investigated for example Followers and Downloads showed a moderate relationship between the two with r=0.5 was found.

Chapter 5

Impact Points and RG Score

There is a need to move beyond traditional metrics such as citations, and paper counts, for evaluating a researcher but citation metrics do not take the author's contribution in consideration [42]. Besides the much criticized two-years window of ISI Journal Impact Factor. With increased use of scholars of the growing social media in the Web 2.0 era, there is a need for new metrics to cope with this change [43]. Researchgate.net came with new features such as the new novel dynamic (alt)metric "Impact Points" and "RG Score". The former Impact Points is a new auto-generated (Alt)metric introduced by Researchgate.net. It builds on ISI Journal impact factor."RG Score" measures an author's impact factor and all his/her activities on Researchgate.net giving an early indicator of his/her publication impact and his/her standing among the Researchgate.net community. It is a dynamic 24/7 evaluation of the researcher activities on Researchgate.net. Since Impact Points on Researchgate.net is based to certain extent on ISI Journal Impact Factor, which is calculated based on average number of citations for a two-year period For example:

A= The current year cites to articles published the last two years by indexed journals.

B=The total number of articles published the previous two years.

Impact factor for year (C) = A/B [44]. This metric is criticized for the two- year window and being one-sided metric where contribution is not considered, ignoring the authors' position on the byline. There is no consensus on accepted rule for contribution of co-authors on a multiple co-authored papers. Traditional metrics do not consider authors contribution in calculating the impact factor. Some scholarly journal like Lancet and NAURE require explicit contribution statement of authors in multi-co-authored papers. There is a clear need for quantitative measures to enable reviewers, evaluation committees and the academic community at large to find out an author's contribution.

5.1 Author Listing on the Byline

There is no single globally accepted method for co-authors listing on the byline, but there are different cultures and Sequence Determine Contribution is the most accepted one. This is why advice given to a researcher to highlight their publications in which they contributed as corresponding authors to notify promotion or tenure committees [45]. There is a viral increase in the number in co-authored publications and a need to show contribution and indicate a paper's credibility [46]. The traditional method was to look at a researcher's achievements through his/her publications in peer-reviewed scholarly journals and the Journal Impact Factor in which he/she published, when it comes to promotion, tenureship or funding. Researchgate.net came with new altmetric tools for evaluating researchers, which is based on (ISI) Journal Impact Factor, and the researches activities in addition to how the scientific community received his/her publications. It is thought that by adding a new dimension to that (alt)metric it will be useful in giving a more meaningful reading. As previously mentioned in calculating the (ISI) Journal Impact Factor it does not consider the authors sequence in its calculations. Here it has been suggested a method for author positing based on Contribution Determines Sequence (CDS) in

calculating the impact points. Impact Points depends partially on (ISI) Journal Impact Factor (JIF) in its calculations, but the author's contribution is disregarded in its calculations. The known author 's positing methods are:

a Alphabetical Order: Authors are positioned alphabetically. In this order an initial letter of authors surname have an effect on the order of authorship. All authors are equally assigned the full impact factor divided by the number of co-authors.

b Sequence Determines Credit: Co-authors are listed according to their contribution. The main author is assigned the full impact factor. The first co-author is assigned the full impact factor minus a "POINT". The second co-author is assigned the full impact factor minus two "POINTS" ...etc.

c First/Last Author Contribution: The first author is accredited the full Impact Factor. The last author is accredited %50. Other authors are credited "POINTS", which are calculated by dividing the impact factor by the total number of authors. In the field of computer science it is an informal practice to sign last for senior researchers, but there is no consensus on the value of other positions.

5.2 Suggested Method

Contribution Determines Sequence (CDS) Method is suggested, in response to the fact of viral increase of multi co-authored publications, calls from scholarly journals publishers and scientific institutions. Assumed was that, the total number of research components are (12) and designed the following scheme on that basis. The total number is divided into four equal divisions (a),(b),(c) and(d)

- Division (a) is assigned 100% of the impact factor.

  First author on top of the division is accredited the full impact factor, the following author is accredited full Impact factor minus a "POINT". The next one is accredited the full impact Factor minus (2) points. (A point is calculated by dividing the division impact factor by 12).

- Division (b) is assigned 75% of the full impact factor.

  First author on top of the division is accredited the full impact factor of the division, and the following author is accredited full impact factor minus a division "POINT". The next one is accredited the full impact factor minus (2) division points. (A point is calculated by dividing the division impact factor by 12).

- Division (c) is assigned 50% of the impact factor.

  First author on top of the division is accredited the full impact factor of the division, and the following author is accredited full impact factor minus a division "POINT". The next one is accredited the full impact factor minus (2) division Points. (A point is calculated by dividing the division impact factor by 12).

- Division (d) is accredited 25% of the impact factor. First author on top of the division is accredited the full impact factor of the division, and the following author is accredited full impact factor minus a division "POINTS". The next one is accredited the full impact factor minus (2) division points. (A point is calculated by dividing the division impact factor by 12).

The Suggested-Method		
A- 100%	1	
Author No.1	1	
Author No.2	0.9167	
Author No.3	0.8334	
B- 75%	0.75	
Author No.4	0.75	
Author No.5	0.6875	
Author No.6	0.625	All other following authors are accredited the 12th. score
C- 50%	0.5	
Author No.7	0.5	
Author No.8	0.458	
Author No.9	0.416	
D- 25%	0.25	
Author No.10	0.25	
Author No.11	0.2292	
Author No.12	0.2084	

5.2.1 CDS Method Features

The method can be used by 8, 12, or 16 authors positions. Decision on that comes from further studies based on the multi-authoring trends in scientific publications. A script was written to crawl Researchgate.net an academic social network, to retrieves data about the co-author's positions on the byline of a research paper. Since there is no formal accepted scheme, the following scheme has been considered to be used on Researchgate.net for more meaningful reading by adding another dimension, "Contribution". Table 5.1: The credit based on Impact factor(14.7) proposed by Teja Tscharntke for PLoS Biology [47].

CDS method is flexible and can be applied in different calculations as can be seen in Table 5.2:

The last position in the scheme is twelve, and each following author will be assigned the same value of number twelve. In Table 5.2, it was moved beyond the scheme of twelve to the scheme of sixteen, to show that the scheme is flexible and functions properly even sixteen positions are considered, and the score is still high.

Table 5.1: Different allocating schemes results proposed by Tscharntke.

Author	SDC	EC	FLAE	PCI	Contribution (%) for PCI	Traditionla Credit
TT	14.7	2.9	14.7	8.8	60	14.7
MEH	7.3	2.9	2.9	2.9	20	14.7
TAR	4.9	2.9	2.9	1.5	10	14.7
VHR	3.7	2.9	2.9	0.7	5	14.7
JK	2.9	2.9	7.4	0.7	5	14.7
SUM	33.5	14.5	30.8	14.6	100	73.5

Table 5.2: Suggested approach (method) for three different reading 8, 12, 16 Co Authors.

Authors Sequence	8-Au.Method	14.7 Reading	12-Au.Method	14.7 Reading	16-Au.Method	14.7 Reading
1st.	1	14.7	1	14.7	1	14.7
2nd.	0.875	12.8625	0.9167	13.475	0.9375	13.78
3rd.	0.75	11.025	0.8334	12.25	0.875	12.86
4th.	0.65625	9.645	0.75	11.025	0.8125	11.94
5th.	0.5	7.35	0.6875	10.105	0.75	11.25
6th.	0.4375	6.43	0.625	9.185	0.703125	10.64
7th.	0.25	3.675	0.5	7.35	0.65625	10.03
8th.	0.021875	3.215	0.458	6.74	0.609375	9.42
9th.			0.416	6.13	0.5	7.35
10th.			0.25	3.675	0.46875	6.8
11th.			0.2292	3.365	0.4375	6.43
12th.			0.2084	3.055	0.40625	5.97
13th.					0.25	3.675
14th.					0.234375	3.375
15th.					0.21875	3.075
16th.					0.203125	2.775

The suggested scheme compared to five different schemes, proposed by different scientists, and the suggested one of twelve positions showed gradual smooth descent with later authors still getting a reasonably good score. A paper of multiple co-authors of (20) with (8.81) journal impact factor, was picked from Researchgate.net as a example, and Credit Allocating Schemes was applied.

5.3 Credit Allocating Schemes

The following five different existing model schemes were applied, in addition to our suggested scheme for comparison:

5.3.1 The Simplest Equalitarian Fractional Allocating

Oppenheim C. [48] proposed a scheme for K authors, where each author received allocation 1/K. This scheme gives an unfair reading to the main author who takes the responsibility, and has done most of the paper's contribution.

Table 5.3: Dividing impact factor by using Fractional Allocating Scheme.

K	1	2	3	4	5	6	7	8	9	10	11	12
12	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73

5.3.2 Tscharntke, Teja Scheme

Teja [48]divided the value of the Impact Factor based on the position of each author (i.e Main author will take the full impact factor IF/r, 1st co-author IF/r, 2 nd. co-author IF/r, ..., K co-authors IF/K), where r=1, 2, 3, 4, 5, ...,K

K: is a number of authors in a paper.

r: is the byline author position.

The scheme was applied to our real data obtaining the following result:

Table 5.4: Dividing impact factor by using Teja Tscharntke Scheme.

K/r	1	2	3	4	5	6	7	8	9	10	11	12
12	8.81	4.41	2.94	2.2	1.76	1.46	1.25	1.1	0.97	0.88	0.8	0.73

This scheme shows a notable difference among authors, if the values of the row (12) were compared. It is notable that there is a gap between the main author and the last co-author, where the main author takes the full value, half one for the first co-author, one third for the second co-author, quarter for the third co-author and so on.....

5.3.3 Arithmetic Allocating Scheme

A well balanced Arithmetic allocating scheme as assumed by the author, proposed by Van Hooydonk [48], for the consequences of the impact of authors. This scheme has been applied to our data as shown in Table 5.5

$$g(r,K) = \frac{2(K+1-r)}{(K*(K+1))}$$

Table 5.5: Dividing impact factor by using Arithmetic Allocating Scheme.

K/r	1	2	3	4	5	6	7	8	9	10	11	12
12	1.35	1.24	1.13	1.02	0.9	0.79	0.68	0.56	0.45	0.34	0.23	0.11

By applying Arithmetic Allocating Scheme, it is clear that co-authors at position five and following got less values than one. Since the most significant row to us was (12), it is very clear and noticeable that the main author will get a very small part of the impact factor value, while he/she has done most of the paper's contribution.

5.3.4 Geometric Allocating Scheme

The following Geometric Allocating Scheme was proposed by (Egghe L., and et al .....) [48], where the main author takes the highest portion of papers' impact factor with

$$g(1,K) = \frac{1}{(2(1-2^{-K}))}$$

while the last co-author takes the smallest(less than one) with

$$g(K, K) = \frac{1}{(2^K - 1)}$$

The Geometric Allocating Scheme is:

$$g(r, K) = \frac{2^{1-r}}{(2 * (1 - 2^{-K}))}$$

Our data was used on the above scheme with the following results as shown in Table 5.6

Table 5.6: Impact factor reading by using Geometric Allocating scheme.

K/r	1	2	3	4	5	6	7	8	9	10	11	12
12	4.64	2.32	1.16	0.58	0.29	0.23	0.19	0.17	0.14	0.12	0.11	0.1

In the above row of Table 5.6, the values of the impact factor are having a sharp decrease and most of the authors get values less than one, which is considered unfair reading.

5.3.5 Tailor Based Allocations(TBA)

Previously mentioned fractional allocations, are attributed to each co-author, based on each contribution, where the summation is equal to one. Serge Galam [48]came with several suggested protocols and all the above fractional allocations are homogeneous, but (TBA) as a suggested name are heterogeneous. The scientist suggested extra bonuses σ for the main author and μ for the last author, where K is the number of co-authors, he starting the formulas from decreasing arithmetic series K, K-1, K-2, ..., 2, 1. The first value K is for the main author and 1 for the last author. The following schemes only work when $K \geq 2$. Which was applied to our impact factor value (8.81).

$$S_k = \frac{(K(1+K))}{2} + \sigma + \mu$$

$$g(1,K) = \frac{(K+\sigma)}{S_k}$$

$$g(K,K) = \frac{(K-1+\mu)}{S_k}$$

$$g(r,K) = \frac{(K-r)}{S_k}$$

g(1,K) and g(K,K) are only defined when $K \geq 2$, while g(r,K) only if $K \geq 3$ with r= 2, 3, ...K-1. The scientist suggested different values to the bonus σ and μ, σ= 2, μ=1, σ=1 , μ=0 and σ= 0, μ=1.

Table 5.7: Impact Factor by using TBA scheme.

K/r	1	2	3	4	5	6	7	8	9	10	11	12
12 $\sigma = 2,\ \mu = 1$	1.52	1.09	0.98	0.87	0.76	0.65	0.54	0.44	0.33	0.22	0.11	1.31
12 $\sigma = 1,\ \mu = 0$	1.45	1.12	1	0.89	0.78	0.67	0.56	0.45	0.33	0.22	0.11	1.23
12 $\sigma = 0,\ \mu = 1$	1.34	1.12	1	0.89	0.78	0.67	0.56	0.45	0.33	0.22	0.11	1.34

In the TBA scheme, the scientist tried to favor the Junior and senior authors in the paper, which is different from all other previous schemes, giving more slots to the main author. By changing the values of bonuses σ, μ, the slot allocated for each author changes as well. It is clear that row (12) shows the main and last author having the same value of the impact factor represented in two functions g(1,K), g(K,K), while the middle values represented in the function g(r,K) decrease. The author still gets less value than one, which might discourage him/her from contributing to multi-authored researches.

5.3.6 Suggested Contribution Determines Sequence (CDS) Method

Impact value of (8.81) was applied in our suggested scheme, where all authors get values based on their contribution (with values greater than one) and close to each other. This will motivate authors to engage in multi-authored researches and give a more acceptable reading. The scheme is applicable in Sequence Determined Credit (SDC) model as shown bellow:

$$g(r,k) = IF - m * (IF/K)$$

only if $r \geq 2$, if r=1 full impact factor is assigned to the main author. The scheme is divided into four groups and points are going to be used according to the following structure:

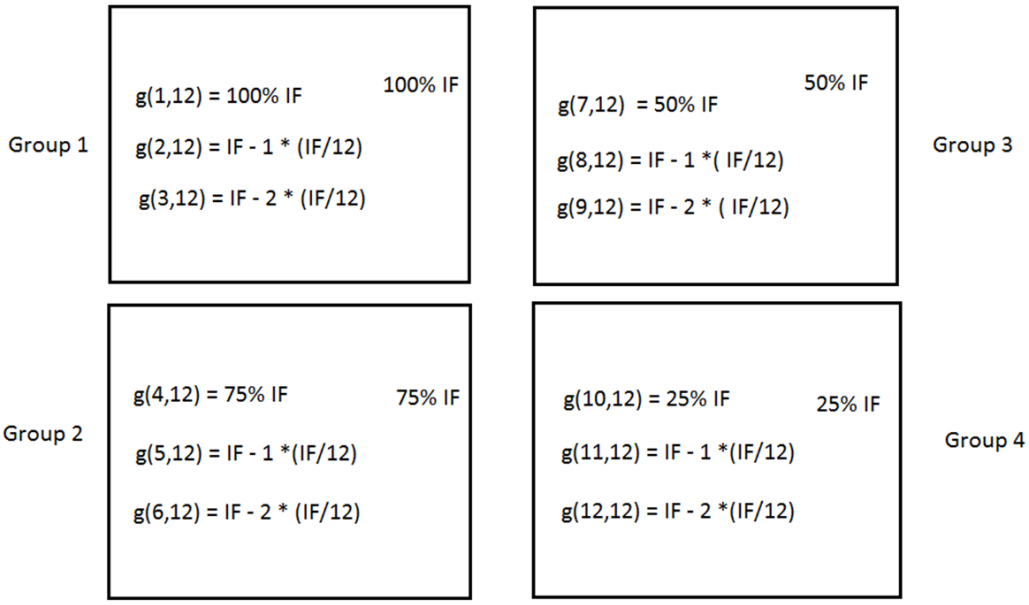


Figure 5.1: Structure of Contribution Determines Sequence (CDS) Scheme.

Where IF = Journal Impact Factor

m= 0, 1, 2.

k= Number of authors inside each group.

K= Total number of authors for a paper.

r= Author's byline position.

The scheme has been implemented in Java language, at the beginning the user was asked to enter the number of authors per paper, and the value of the impact factor to start the process. The following algorithms show the scripts:

5.4 CDS Calculation Scripts

An array was created to save all the generated results of the impact factors (IF) and whatever was the number of authors per paper, this number will be divided into four groups. For any odd number that is not dividing by 4, it will be assign automatically to the last group. The following Figure 5.2 indicates the calculation of Group 1, with same concept group 2, 3, 4 generated. Any one comes after the 12th. position. He/she will take the same value of 12.

```
1: Start program
2: Declare a, scanner=newScanner(system.in) =
   getting initial input
3: Declare int NumofAuthors,double initial_IF
   getting initial_IF and Number of authors
4: Declare double IF,IF_values[NumofAuthors]
5: Assign int count=0, k=12, int iterator, int temp=NumofAuthors, a=NumofAuthors/4
   helper variables used in code to solve the logic
6: Assign IF=Initial_IF
7: for (iterator=0, iterator greater than a, increment iterator)
8: result =0
9: result=IF-(iterator*(IF/k))

10: IF_values[count++] = result

11: −− temp by 1
12: End for-loop
13: End
```

Figure 5.2: Pseudo Code for Calculating The IF for Group1.

```
1: Start program
2: Assign IF=Initial_IF * 3/4
3: for (iterator=0, iterator greater than a, increment iterator)
4: result =0
5: result=IF-(iterator*(IF/k))

6: IF_values[count++] = result

7: −− temp by 1
8: End for-loop
9: End
```

Figure 5.3: Pseudo Code for Calculating The IF for Group2.

1: Start program
2: Assign IF=$Initial\_IF * 2/4$
3: **for** (iterator=0, iterator greater than a, increment iterator)
4: result =0
5: result=IF-(iterator*(IF/k))

6: $IF\_values[count++] = result$

7: $--$ temp by 1 **end** try block
8: **End** for-loop
9: **End**

Figure 5.4: Pseudo Code for Calculating The IF for Group3.

1: Start program
2: Assign IF=$Initial\_IF * 1/4$
3: **for** (iterator=0, iterator greater than a, increment iterator)
4: result =0
5: result=IF-(iterator*(IF/k))

6: $IF\_values[count++] = result$

7: $--$ temp by 1 **end** try block
8: **End** for-loop
9: **End**

Figure 5.5: Pseudo Code for Calculating The IF for Group4.

1: Start program
2: Declare int tempcount=count -1
3: **While** (temp greater than or equal 1) **do**
4: IF_values[count++]=IF_values[tempcount]

5: $--$ temp by 1 **end** try block
6: print output values
7: **for** (iterator=0, iterator less than NumofAuthors, increment iterator) **do** Execute iterator get and send response
8: System.out.println(IF_values[iterator])

9: **End** for-loop
10: **End**
11: Close program

Figure 5.6: Pseudo Code for Calculating The IF for Group5.

When the inputs are K =12, IF = 8.81, The outputs from this code are shows in the following table 5.8:

Table 5.8: Results of Suggested Contribution Determines Sequence Scheme

K/r	1	2	3	4	5	6	7	8	9	10	11	12
12	8.81	8.08	7.34	6.61	6.06	5.51	4.41	4.04	3.67	2.2	2.02	1.84

In Table 5.8: results proved the value of the suggested scheme, improving the reading of each author by assigning the value he/she deserved according to the scheme. Since the values are very close to each other, and even sixteen ranking is applied, the scheme still functions. Our scheme was compared with others and the following figure shows the difference.

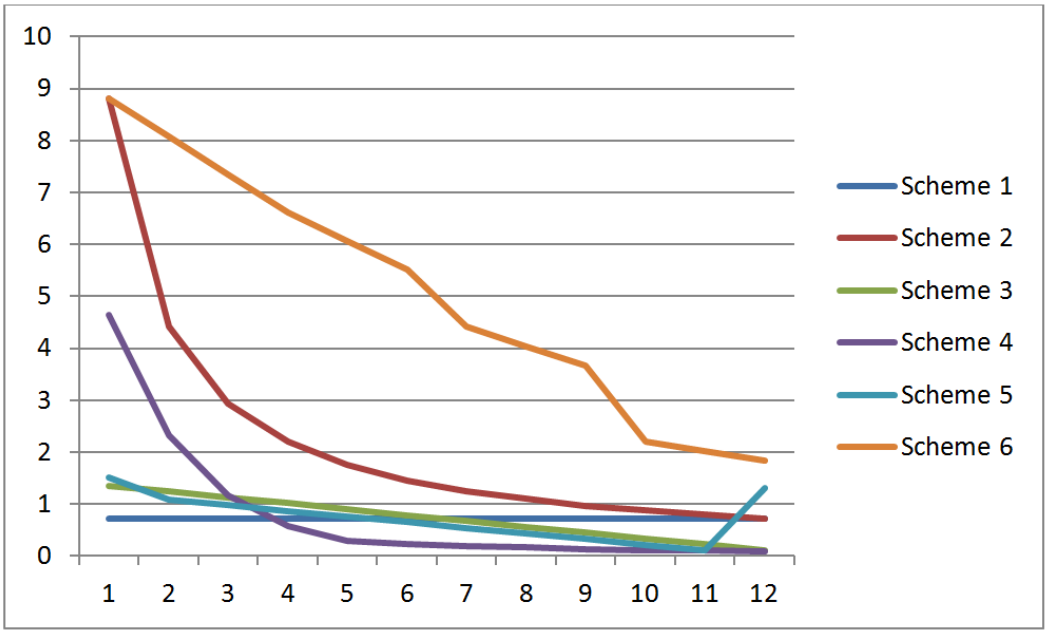


Figure 5.7: Suggested Contribution Determines Sequence Compared to Other Schemes.

Our scheme in Figure 5.7, is almost linearly modeled, as clear smooth descending order, while scheme 2 and scheme 4 show exponentially descending with a gap between any two authors. Our scheme deleted this gaps.

5.5 Experiment

Areal data was applied to our scheme model, to check functionality. A sample of two researchers on Researchgate was picked, taking in consideration that they almost have the same number of publications and the value of impact points on their profiles. Jointly co-authored articles were identified and checked their author position on the byline.Two schemes (scheme 2,scheme 4) in addition to our suggested one were applied, which are the closest ones to our scheme as shown in Figure 5.7 scheme2, schem4 in addition to our scheme. Summations of 1st. author and 2nd. author in each scheme were found and it showed that 1st. author comes constantly in higher rank than 2nd. author. The sum was divided, sum (author1)/sum (author2) to get a factor, then the factor was compared to another factor obtained from the impact points (author1)/impact points (author2). The same approach was applied to scheme4 and scheme6 and the following results showing the difference: impact points(author1)/impact points(author2) = 1.5

scheme 2 = 2.2

scheme 4 = 2.8

scheme 6 (suggested scheme) = 1.2

Results shows the (suggested scheme) gives a closer reading than other schemes to the factor obtained from the impact points on Researchgate.net. In this case the gaps between researchers have been eliminated.

The Figure 5.8, shows an illustrative example of how the suggested CDS Method can be applied in a contribution guidelines to be used by multiple co-author team to determine contribution. We all agree that the best to allocate contribution is the researcher himself/herself.

The CDS method was used to improve the reading of RG score and impact factor. Five records were picked from the sample to give example as in Table 5.9, to give more meaningful

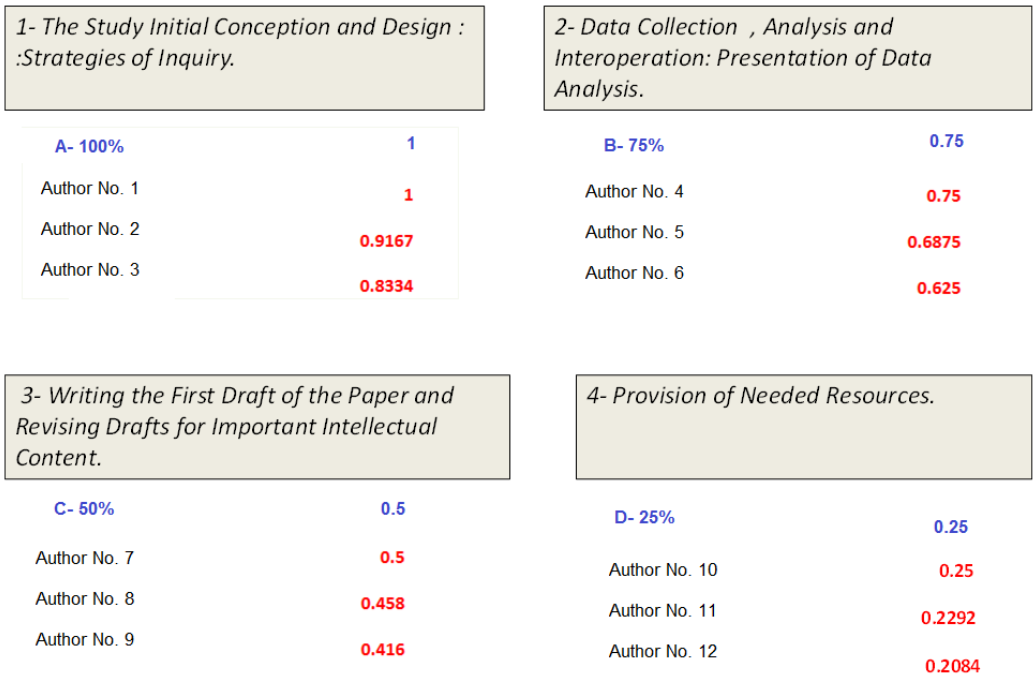


Figure 5.8: Illustrative Basic Research Components on Contribution Determines Sequence (CDS) Method.

reading compared to Researchgate.net one.

(PB) Researcher has different positions in all his publication, where he came (188) times as the main author out of (247) publications. The assumed journal impact factor =1 for the purpose of this example. On Researchgate.net he will be credited the full (247) journal impact factor, but calculated according to suggested scheme his score will drop to (240.17). The same is applicable to (IE), where he came (132) times as the first co-author out of total of (315) publications. By Researchgate.net he will be credited the full impact factor for all his (315) publications, but when applying the suggested scheme his score drops to (282). (AE) Researcher has different positions in all his publication, where came (198) times as the second co-author out of total of (458) publications. By Researchgate.net he will be credited the full impact factor for all his (458) publications, but when the suggested scheme is applied, his score

Table 5.9: Suggested Contribution Determines Sequence Method is applied to five different researchers with different positions. (Ranking first initials were used for privacy issues)

Author	M.A	1st C.A	2nd C.A	3th C.A	4th C.A	5th C.A	6th C.A	RG Reading	CDS Scheme Reading
PB	188/188	40/36.67	15/12.50	4/3				247	240.17
IE	75/75	132/121.00	72/60.00	22/16.5	12/8.25	2/1.25		315	282
AE	37/37	108/99.00	198/165.01	78/58.5	26/17.88	9/5.63	2/1	458	384.02
EB	24/24	22/20.17	4/3.33	26/19.5	7/4.81	2/1.25		85	73.06
DW	2/2	0	1/0.833	0	3/2.06			6	4.89

drops to (384.02). (EB) Researcher has different positions in all his publication, where came (26) times as the third co-author out of total of (85) publications. By Researchgate.net he will be credited the full impact factor for all his (85) publications, but when the suggested scheme is applied, his score drops to (73.06). The same can be said about (DW) Researcher who has different positions in all his publications, where he came (3) times as the fourth co-author out of total of (6) publications. By Researchgate.net he will be credited the full impact factor for all his (6) publications, but when the suggested scheme is applied, his score drops to (4.89). Applied were the other existing schemes and the results were compared to our suggested one. Table 5.10 represents how CDS scheme (the suggested one) gives more meaningful reading and much more closer to the Researchgate.net reading than other schemes.

Table 5.10: Comparison among schemes reading and finding the closest one to Researchgate.net reading.

	Researchgate.net Reading	S.E.F.A Scheme	T.T Scheme	A.A Scheme	G.A Scheme	TBA Scheme	CDS Scheme
PB	247	20.58	213.95	36.19	112.25	35.45	240.17
IE	315	26.22	172.98	42.89	85.3	39.75	282
AE	458	38.11	182.76	58.54	79.65	52.29	384.02
EB	85	7.06	44.54	10.94	20.93	10.3	73.06
DW	6	1.25	2.93	0.73	1.28	0.68	4.89

5.6 CDS Main Advantages

The main advantages of the new modified, supplemented impact scores are:

A - Defending the contribution of junior researchers.

B - Giving more indicative pointer to the academic activities, not tied solely to published researches Impact Factor only.

C - Creating more awareness among researchers of the importance of authorship since it is calculated in impact score towards academic performance measurement and reputation building.

D - Highlighting the need to develop a guide for a standard, accepted author positioning system. The known metrics disregard "Contribution" as a dimension in measuring academic performance.

E - Improving the readings of "Impact Points" and "RG Score" on Researechgate.net.

F - A research-oriented contribution allocating method with smooth descending within each group. Exuberant literature is available in the field of Computer Science. It is much advised that Canadian researchers in the field of Computer Science sit down and determine each researcher contribution before starting a research project. This is not easy to do since sometimes an advise might contribute more than writing lines in a research paper. There was hope that the scheme might be helpful as well in designing accepted "contribution guidelines Guide" which might be included in the university manual of style.

5.7 Summary

Researchgate.net came with the new Impact factor calculation based on ISI Journal Impact Factor to highlight the quality of contribution a researcher makes. A new method has been sug-

gested for calculating contribution when calculating the"Impact Points" on Researchgate.net. An illustrative model has been presented for using our suggested Contribution Determines Sequence method (CDS), which showed good results when compared to other credit allocation schemes. Our suggested method involves the author in determining his/her contribution not relying totally on followed calculation practiced presently. Aiming at putting the author behind the steering wheel since he/she is the best to determines his/her contribution.

Chapter 6

Conclusion and Future Works

Generally there are different types of web crawlers, general and focus crawlers, used for different goals, with different crawling techniques. Crawler developed with the development of the Internet and the challenge of huge data on the Internet and interactivity that led to the development of deep Internet and Rich Internet Applications (RIA) crawlers. Presently there are many types of crawlers to meet the constant change on the Internet. Incremental Crawler is a traditional one which to refresh its collection, replaces the old documents with newly downloaded ones. The advantage of incremental crawler is that the user is provided with valuable data achieving data enrichment and maintaining network bandwidth. A Focused Crawler or topical crawler, downloads related pages determining way forward relevancy. It is economical on hardware and the network resources. A Distributed crawler applies distributed computing techniques for extensive web coverage using Page Rank algorithm. The main advantage of this crawler is flexibility. A Parallel Crawler depends on page freshness and page selection allowing for multiple crawling by running many crawlers in parallel. Development of a suitable or effective crawler requires taking a number of challenges that subtly interfere and create issues, especially in large scale web crawlers. As mentioned, there are numerous challenges, however,

some to mention in this context are, politeness for the web servers, duplicate detection, URL normalization, queue maintenance of un-fetched web pages, re-crawling as well as to prevent spider traps. On the other hand, in case of large scale crawlers, throughput increment and resource utilization are the main issues that have to be managed in order to liberate coverage. Our crawler was written in Java language using different software and libraries. Another different software were used such as MS Excel, to find the correlation coefficient and Minitab17 to find ANOVA table. The crawler retrieved real data which was analyzed to highlight the performance of Canadian researchers in the field of computer science on Researchgate.net. It came with new features such as the new novel dynamic (alt)metric " Impact Points" and "RG Score ". The former is Impact Points is a new auto-generated (Alt)metric introduced by Researchgate.net. It builds on ISI Journal impact factor. RG Score measures an authors impact factor and all his activities on Researchgate.net giving an early indicator of his/her publication impact and his/her standing among the Reasearchgate.net community. Our contribution is development of a crawler and Contribution Determines Sequence (CDS) method with required scripts, which gave better results compared to other credit allocation methods. To test the crawler, it was run on the academic social network, Researchgate.net from April 3-June 28 2014 and real data was retrieved. The retrieved data was analyzed to highlight the performance of Canadian Researchers, in the field of Computer Science on Researchgate.net. Data analysis was done from the collaboration and (Alt)metrics perspectives. Data analysis highlighted the Canadian researchers performance of Researchgate.net in the field of Computer Science showed the presence of correlation between a student's output and the number of co-authored papers published with his/her supervisor. Co-authored publications of supervisor/student were identified, Post Doc. came on the top in information seeking and knowledge sharing activities to confirm other researches done on Academic.edu. Investigation of Researchgate.net metrics indicated a corre-

lation between Views and Download and statistics revealed that academic ranking has no effect on Downloads.

In the future work the plan is to develop the crawler to crawl multiple sides in a parallel. In addition to conduct more in-depth analysis of the behavior of the observed group and to extend the study to different disciplines and geographical areas, to make comparison of researchers from different disciplines. A multidisciplinary study, might develop and suggest practical applicable method for determining author's contribution with required related programs. The suggested credit allocating method (CDS) in this research, can be developed on the light of future studies to improve "Impact Points" calculation on Researchgate.net. When the developed (CDS) is accepted by the scientific community, it is advised that Canadian researchers in the field of Computer Science sit down and determine each researcher contribution before starting a research project, then to be reviewed at later stages. Determining contribution is not easy to do since sometimes an advise might contribute more than writing lines in a research paper, but the suggested method will, hopefully allow that. There is exuberant literature in the field of Computer Science and the illustrative example using the suggested CDS method might trigger further investigations into designing accepted "Contribution Allocation Guide", which involves the researcher decision not mere automatic calculation. This guide can be included in the university Manual of Style to be used by authors in multiple co-authored researches.

Appendix

Crawler's Algorithms

In the appendix we listed our crawler's algorithms based on their sequence of creation. All our algorithms in addition to these mentioned in Chapter 3 are listed under the following categories :

1. Retrieving data about researchers in more details starts from algorithm (1-7)
2. Crawling information about publications starts from algorithm (8-11)
3. Crawling information on questions/ answers starts from algorithm (12-16)

Algorithm 1: Crawling Information about Researchers

Data: This Algorithm allows the connection to departments and institutions in Researchgate once we are logging in to the website.

Result: Expecting to retrieve data searched according to institutions and departments which in this case are computer science departments only for all Canadian Universities or institutions are participating to this Academic Social Network.

```
1  Start program;
2  Login researchgate;
3  duin.fetchIdNameOfInstitute(httpclient);
4  getting institute name;
5  for (String keyinstituteName :map.keySet()) do
6      iterate keyinstituteName;
7      List<String> departments = newArrayList<String>();
8      departments.addAll((map.get(keyinstituteName)));
9      get department add to list;
10     for (String department : departments) do
11         if (offset == 0) then
12             get url
13         else
14             add offset to url
15         Crawl url and get response;
16         if (response1 != null) then
17             Source source = new Source(entity.getContent());
18             string = source.toString()
19         if (!string.isEmpty()) then
20             get jsonobjectsize
21         if (sizeofjsonobject > 0) then
22             increment startindex by 1;
23             increment offset by 1;
24             (startindex != sizeofjsonobject);
25             Catch exception;
26             Execute finally block;
27 Fetch researcher details;
28 Close program
```

Algorithm 2: Client Form Login

Data: Using httpGet to enter login credentials onto research login from that will be returned and post to send the data back to researchgate for authentication.

Result: Expecting to be logged in successfully after login credentials have been authenticated by researchgate.net.

```
1  Start program;
2  HttpGet httpget = new HttpGet("https://www.researchgate.net/");
3  Get login page from https://www.researchgate.net/;
4  CloseableHttpResponseresponse1 = httpclient.execute(httpget);
5  Execute httpget and send response;
6  begin try
7      HttpEntityentity = response1.getEntity();
8      Print response;
9      Release all resources httpclient
11 begin finally
12     response1.close()
13 Send login credentials through post request ;
14 Repeat step 11;
15 List<Cookie> cookies = cookieStore.getCookies();
16 Get cookies from cookieStore and add to arrayList;
17 if (cookies.isEmpty()) then
18     display none
19 else
20     begin cookie-loop
21         iterate for loop and display all cookies
22 Close response;
23 Return httpclient;
24 End;
```

Algorithm 3: FetchUniversityId_Name

Data: To get accurate response, after logging in I was able to modify my query to be able to fetch only universities from Canada

Result: Expecting data to return only universities in Canada.

```
Start program;
Declare Listofelements = null, offset = 0, testurl = null;
Create List of institutenames;
Create instance of institutenames class;
institutenames = rf.listofInstitute();
Create arrayList for universityId,universityNames,keyuniversityNames;
Assign http://www.researchgate.net/institutions/Canada?order=rgScore&method=total
to testurl;
if (testurl != null) then
    testurl = testurl + offset;
    get the source code testurl
get listofelements by class name;
ListOfelements = doc.getAllElementsByClass("lfname");
iterate lisofelements using for-each loop;
for (Element eleee : ListOfelements) do
    List<Element> eless = eleee.getAllElements("a")
if (!eless.isEmpty()) then
    get first attribute in list
attribute split and replace;
String[] ss = attss.toString().split("/");
String names = ss[1].replace("n"", "");
if (institutenames.contains(names)) then
    add names to keyuniversityNames,universityNames
create insance for attribute class;
get attribute by id;
String str = atts.getValue("id");
if (str != null) then
    split str ;
    add universityId;
offset = offset + 1;
add universityId, universityNames, keyuniversityNames to map;
fetch department;
fetchpdeparments.FetchDepartment(map, httpclient);
End
```

Algorithm 4: InstituteNames

Data: Query written to read and return institutions in Canada.
Result: All Institutions in Canada were returned as a response to my query.

```
Start Program;
Create file class;
Assign BufferedReader reader = null;
Create listofinstitute as a arraylist;
begin try
    Read file;
    reader = new BufferedReader(newFileReader(file))
assign text = null;
while ((text = reader.readLine()) != null) do
    index = text.lastIndexOf('/');
    if (index > 0) then
        String strtr = text.substring(index + 1);
        Add strtr to listofinstitute
Catch exception;
Execute finally block;
Read file;
Return listofinstitute;
End
```

Algorithm 5: FetchDepartments

Data: University ids are iterated using for each loop to return researcher data from computer science departments.

Result: Expected to return names of researchers from computer science departments in the Canadian Universities and Institutions on researchgate.net

```
1  start program;
2  assign j = 0, passurl = null, universityName = null, keyuniversityName = null,
   departmentName = null, string = null, responsejson = null;
3  create arraylist for keyuniversityNames, universityids, universityNames;
4  get keyuniversityNames, universityids, universityNames from map;
5  add all values to arrayList i.e;
6  universityids.addAll((map.get("UniversityIds")));
7  universitynames.addAll((map.get("UniversityNames")));
8  keyuniversityNames.addAll((map.get("KeyUniversityNames")));
   /* iterate universityids using for-each loop                              */
9  for (String universityid : universityids) do
10 |  create departments as arraylist;
11 |  universityName = universitynames.get(j) ;
12 |  keyuniversityName = keyuniversityNames.get(j)
13 declare passurl;;
14 add universityName, keyuniversityName to pass url;
15 assign
   "https://www.researchgate.net/signup.SignUp.ajaxDepartments.html?query=&institutionId=";

   + universityid + "&institutionName=" + universityName to passurl;
16 crawl passurl;
17 get response;
18 get entity from response;
19 create jsonobj;
20 if (lengthofjsonobject > 0) then
21 |  while (lengthofjsonobject != startindex) do
22 |  |  get departmentname;
23 |  |  for (int ii = 0, ii < ss.length, increment ii) do
24 |  |  |  get computerscience department;
25 |  |  |  increment ii by 1
26 |  |  increment startindex by 1;
27 add departmentnames to map;
28 call researchercrawler;
29 fetch researchername;
30 end
```

Algorithm 6: FetchResearcherDetails

Data: Using for loop, I intend to get researcher details.
Result: Expected to return researcher profile details like departments, number of impact points, citations, total number of downloads of researcher's publications and others.

```
start program
Assign http://www.researchgate.net/profile/ to profileUrlPrefix
declare xpath
/* iterate researchernames using for-each loop                    */
for (String name : Names) do
    profileUrlPrefix = profileUrlPrefix + name
create instance Researcher
if (profileUrlPrefix != null) then
    declare response4 = null, source = nullendif
    for (int i = 0; i < xpath.length, increment i) do
        if (i = 0) then
            source = (Source) navigator.getDocument(profileUrlPrefix)
        if (i = 1) then
            (researcher).setInstitute
        if (i = 2) then
            (researcher).setDepartment
        if (i = 3) then
            (researcher).setFollowers
        if (i = 4) then
            (researcher).setPublication
        if (i = 5) then
            (researcher).setView
        if (i = 6) then
            (researcher).setDownloads
        if (i = 7) then
            (researcher).setCitations
        if (i = 8) then
            (researcher).setImpactpoints
        if (i = 9) then
            HttpUriRequest search1 = RequestBuilder.get().setUri(new
            URI("https://www.researchgate.net/profile/"+ name+ "/contributions/?ev=brs_act")).build()
            Goto contributions page
            Crawl source code
            (researcher).setQuestions
        if (i = 10) then
            (researcher).setAnswers
    if (researcher != null && researcher.getName() != null) then
        set all details to researcher
```

Algorithm 7: Details

Data: After details have been collected in algorithm 6, results are appended to and stored in a document format.

Result: Expected to append information into a document which in this case I used .csv format to append into excel spreadsheet for the ability to analyse and calculate statistics of data returned.

```
1  Start program;
2  assign csv_seperator = ",";
3  assign i = 1;
4  BufferedWriterbw = newBufferedWriter("filepath");
5  StringBuffer oneLine = new StringBuffer();
6  Write all details to file path using stringBuffer;
7  Write sno,name,institute;
8  Write department,followers,publications;
9  Write views,downloads,citations;
10 Write impactpoints,no.of questions,no.of answers;
11 Add to file;
12 goto new line;
13 bw.newLine();
14 for (Researcher researcher : researchers) do
       /* iterate all researchers                                    */
15     StringBuffer oneLine1 = new StringBuffer();
16     oneLine1.append( researchername);
17     oneLine1.append(institutename);
18     oneLine1.append(department);
19     oneLine1.append(followers);
20     oneLine1.append(publications);
21     oneLine1.append(views);
22     oneLine1.append(downloads);
23     oneLine1.append(citations);
24     oneLine1.append(impactpoints);
25     oneLine1.append(no.of questions);
26     oneLine1.append(no.of answers);
27     add to file;
28     newline();
29     increment i by 1
30 flush();
31 close();
32 catch exceptions;
33 end
```

Algorithm 8: PublicationCrawler

Data: To get publication details of each researcher on researchgate.
Result: To return publication details of reseachers from computer science departments in the Canadian institutions and Universities.

1 Start program;
2 Create instance for BasicCookieStore class;
3 Build httpclient;
4 Create instance for clientformlogin class;
5 $ClientFormLogin\ cfl = new\ ClientFormLogin()$;
6 Login researchgate;
7 Get httpclient;
8 call FetchUniversityIdName;
9 Fetch universityid;
10 End

Algorithm 9 : PublicationUrl

Data: Using publication details class we are able to retrive publication url based on country, state university and department where researcher's work was published.

Result: Return publication details based on department and universities specified.

```
1  start program
2  declare source = null, nextpageid = null, url = null, departmentId = 0,size
3  store publication url to publication_names list
   List<String> publication_names = new ArrayList<String>()
4  iterate publicationurl using for-each loop
5  for (String keyinstituteName : map.keySet()) do
6      get keyinstituteName & add to department list
7  iterate departments using for-each loop
8  for (String department : departments) do
9      while (nextpageid = null) do
10         if (nextpageid == null) then
11             assign
               https://www.researchgate.net/publicinstitutions.DepartmentContributionsContent.html?institutionKey="+
               keyinstituteNam+ "&departmentKey="+ department to url
12         else
13             add nextpageid, departmentid to url
14         crawl url
15         get response
16         begin try
17             entity = response.getentity()
18         get source
19         convert source to jsonobject
20         assign jsonobject id = null
21         if (nextpageId == null) then
22             jsonarray = id.getJSONArray("feedItems")
23         if (departmentId == 0) then
24             getid from jsonarray
25             Nextpageid = id.get("olderthan")
26         else
27             get jsonarray
28         declare i = 0;
29         if (jsonarray.length() ! = 0) then
30             begin try
31                 while (i ! = size) do
32                     get jsonresponse
33                     Declare sofi=jsonresponse.length()
34                     Declare j = 0
35                     while (j ! = sofi) do
36                         get publicationurl
37                         if (!publication_names.contains(publicationurl)) then
38                             add publicationurl
39                         increment j by 1
40                     Increment i by 1
41         Catch exception
42 Call publicationdetails class
43 End
```

Algorithm 10: PublicationDetails

Result: Returns data from algorithm 9.

Start program
Declare $publicationurl = null, refer = null$
Iterate publicationtitle using for-each loop
for *(String url : publicationtitles)* **do**
    assign "https://www.researchgate.net/" + url to publicationurl
    Create authors arrayList
    Declare publicationdate=null
    get source content
    if $(publicationurl != null)$ **then**
        create object for each publication
    getAllElementsByClass("subheader-small-support has-right-col")
    all elements add to listofelement
    iterate listofelement using for-each loop
    for *(Element sas : ListOfelement)* **do**
        $lisofelements = sas.geelementbyclass(\text{ic} - content\text{j})$
    iterate lisofelements
    for *(Element eleee : ListOfelements)* **do**
        get element from 'h1'
    begin try
        set publicationname
    Catch exception
    $ele = eleee.getAllElements("div")$
    if $(!ele.isEmpty())$ **then**
        getAllElementsByClass("js-expander-container js-expander-collapsed")
    iterate listofelements
    set authorsnames to publicationdetails
    getAllElementsByClass("pub-details")
    set publicationdate
    for *(String datee1 : dates)* **do**
        if *(datee1.contains("/"))* **then**
            split("/")
        check month & year length
        set published date
        $ListOfelements = ele.get(0).getAllElementsByClass("action - container")$
        Iterate lisofelements use for-each loop
        Set view for publicationdetails
    $ListOfelements = ListOfelement.get(0).getAllElementsByClass("c - col - right")$
    repeat step 35
    get all elements by strong tag
    set downloads for publicationdetails

Algorithm 11: PublicationDetailsCSV

Result: Returns and appends publication details into CSV.

```
start program;
Assign csv_separator = ",", i = 1;
BufferedWriterbw = newBufferedWriter("filepath");
StringBuffer oneLine = new StringBuffer();
Write sno,publication name, institute Name;
Write views, downloads, date of publication;
Write main author, co-author;
Write 20 co-author names;
bw.newLine();
iterate publicationdetails using for-each loop;
for (ResearcherPublications researcherpublication : publicationdetails) do
    StringBuffer oneLine1 = new StringBuffer();
    onLine1.append("PublicationName");
    oneLine1.append(CSV_SEPARATOR);
    onLine1.append(institutename);
    oneLine1.append(CSV_SEPARATOR);
    if (researcherpublication.getViews() == null) then
        setview empty
    onLine1.append(views);
    onLine1.append("downloads");
    if (researcherpublication.getDateOfPublishing() == null) then
        set dateofpublishing empty
    Online1.append(dateofpublishing);
    online1.append(csv_separator);
    List<String> names = researcherpublication.getAuthornames();
    Declare size = names.size();
    Declare name, j = 0;
    while (j != size and j ≤ 20) do
        name = names.get(j);
        online1.append(name);
        online1.append(csv_separator);
        increment j by 1
    for (int k = j, k ≤ 20, increment k) do
        online1.append(csv_separator)
    write new line;
    increment i by 1
release all resources;
close connection;
catch all exceptions;
end
```

Algorithm 12: QuestionAnswerCrawler

Data: To get all questions asked and answered by researchers and answers provided to their questions.

Result: Expected to show how researchers are interacting with one another on researchgate through the questions they ask and answers that are provided to them and to establish relationship between them.

```
1  Start Program;
2  Create instance for BasicCookieStore;
3  Build httpclient;
4  Login researchergate & get httpclient;
   httpclient = cfl.fetchhttpclient(httpclient, cookieStore);
5  Fetch institute id, name and keyinstitutename;
   duin.fetchIdNameOfInstitute(httpclient);
6  Create arrayList for names;
   List<String> Names = new ArrayList<String>();
7  Assign offset = 0, sizeofjsonobject = 0, startindex = 0, string = null, url = null;
8  iterate keyinstituteName;
9  List<String> departments = new ArrayList<String>();
10 departments.addAll((map.get(keyinstituteName)));
11 get department add to list;
12 for (String department : departments) do
13     while do
14         if (offset == 0) then
15             get url
16         else
17             add offset to url
18         Crawl url and get response;
19         if (response1 != null) then
20             Source source = new Source(entity.getContent());
21             string = source.toString()
22         if (!string.isEmpty()) then
23             get jsonobjectsize
24         if (sizeofjsonobject > 0) then
25             while (startindex != sizeofjsonobject) do
26                 increment startindex by 1;
27                 increment offset by 1
28             Catch exception;
29             Execute finally block
30 Fetch researcher details;
31 Call FilterNameHavingQuestion class;
32 End
```

Algorithm 13: FilterNameHavingQuestion

```
   Data: To filter questions from answers.
   Result: Expected to return questions asked by researchers.
 1 Start Program;
 2 Create arrayList for researcherName;
 3 Assign http://www.researchgate.net/profile/ to profileUrlPrefix;
 4 Declare xpath;
 5 Assign profileUrlPrefix.lastIndexOf('/') to index;
 6 iterate researchernames using for-each loop;
 7 for (String name : Names) do
 8     profileUrlPrefix = profileUrlPrefix + name
 9 if (profileUrlPrefix != null) then
10     assign result = null, response4 = null, source = null;
11     create instance for navigator begin try
12         set uri;
13         crawl uri & get source;
14         source = new Source(entity3.getContent());
15         create instance for xpath;
16         XPath expr = new JerichoXPath(xpath, navigator);
17         result = expr.evaluate(source)
18     catch exception;
19     execute finally block;
20     if (xpath != null) then
21         if (result instanceof Element) then
22             print element & element content
23         else if (result instanceof List) then
24             typecast result into list
25         if (!elements.isEmpty()) then
26             add name to researchername
27         else
28             display null
29 profileUrlPrefix = profileUrlPrefix.replace(toBeReplaced, "");
30 call FetchResearcherDetailsQA;
31 end
```

Algorithm 14: FetchResearcherDetailsQA

Data: To find the path from which questions are coming from and sorting them according to researchers from Computer science departments in the Canadian Universities.

Result: Expected to return questions asked by researchers with their names and details of profile on researchgate.net.

```
1  start program;
2  Assign http://www.researchgate.net/profile/ to profileUrlPrefix;
3  declare xpath;
4  iterate researchernames using for-each loop;
5  for (String name : Names) do
6      profileUrlPrefix = profileUrlPrefix + name
7  create instance for ResearcherHavingQuestionDetail;
8  if (profileUrlPrefix != null) then
9      declare response4 = null, source = null
10 for (int i = 0; i < xpath.length, increment i) do
11     if (i = 0) then
12         getDocument(profileUrlPrefix) assign to source
13     create instance for xpath;
14     result = expr.evaluate(source);
15     if (xpath[i] != null) then
16         if (result instanceof Element) then
17             print element & element content
18         else if (result instanceof List) then
19             typecast result into list;
20             for (int j = 0, j < elements.size(), increment j) do
21                 element conten;
22                 execute switch case;
23                 if (i = 0) then
24                     (researcher).setResearcherName
25                 if (i = 1) then
26                     (researcher).setUniversity
27                 if (i = 2) then
28                     (researcher).setDepartment
29 print result instaceof result, number,Boolean value;
30 if (researcher != null && researcher.getResearcherName() != null) then
31     put name,researcher into map
32 profileUrlPrefix = profileUrlPrefix.replace(toBeReplaced, "");
33 call QuestionAndAnswerDetails;
34 end
```

Algorithm 15: QuestionAndAnswerDetails

Data: To fetch answers submitted to questions asked by researchers although some answers were provided by others from other faculties or departments the filter was able to sort through it and provide all answers irrespective of where its coming from.

Result: Expected to return each answer to each question asked by each researcher who provided an answer.

```
Start Program
Declare noofpage = 0, noofanswer = 0
Declare researcherquestion = null, baseUrl = null, questionurl = null, questionandanswer = null,
url = null,page = null,resarhername = null,department = null,institute = null,source =
null,baseurl = null
for (String name : map.keySet()) do
    assign "https://www.researchgate.net/profile.ProfileContributionsContent.html?account_key="+ name to url
get response
if (response != null) then
    get source
    convert source to jsonobject
    for (int i = 3, i ≤ 3; increment i) do
        form baseurl
        while (page != null) do
            if (page != null) then
                split baseurl
                questionandanswer = "https://www.researchgate.net/"+baseurl[0]+"page="+noofpage
            else
                assign https://www.researchgate.net/ + baseUrl to questionandanswer
            get response
            get source from content
            if (response != null) then
                convert source to jsonobject create jsonarray for questions begin try
                    noofpage = jsonobj.getInt("page") + 1
                catch exceptions
                assign length of questions to no.ofquestion
                if (noofquestion != null) then
                    declare question = 0 while (noofquestion != questionno) do
                        get questionurl from jsonobject
                        append questionurl to url
                        get response from url
                        create map for researcheranswerdetails
                        if (response != null) then
                            get source from entity
                        listofelements = source.getAllElementsByClass("c − content")
                        listofelement = listofelements.getAllElementsByClass("topic − post − title")
                        if (!listOfelement.isEmpty()) then
                            iterate listofelement by using for-each loop
                            for (Element element : listOfelement) do
                                get researcherquestion
                        else
                            set researcherquestion empty
```

Algorithm 16: QuestionAnswerDetailsCSV

Data: To append information obtained from Algorithm 10 to 15 into CSV file
Result: Expected to return all data and append into CSV file for easy assessment.

```
start program;
Assign csv_separator = ",", i = 1;
BufferedWriterbw = newBufferedWriter("filepath");
StringBuffer oneLine = new StringBuffer();
Write s.no,Researcher Name,University,Department;
Write Question,Answer;
for (int j = 1, j ≤ 20, increment j) do
    write 20 answers
write new line;
bw.newline();
for (ResearcherHavingQuestionDetail researcherquestionanswer : researchers) do
    create map for answer;
    create set for question;
    all questions add to arrayList;
    get answer from map add to list;
    StringBuffer oneLine1 = new StringBuffer();
    oneLine1.append(i);
    oneLine1.append(CSV_SEPARATOR);
    oneLine1.append.append(ResearcherName);
    oneLine1.append(CSV_SEPARATOR);
    oneLine1.append.append(University);
    oneLine1.append(CSV_SEPARATOR);
    oneLine1.append(department);
    oneLine1.append(CSV_SEPARATOR);
    oneLine1.append(question);
    oneLine1.append(CSV_SEPARATOR);
    declare size= answerdetails.size();
    declare j = 0;
    while (j != size && j ≤ 20) do
        create instance for ResearcherAnswerDetails;
        ResearcherAnswerDetails answerdetail = new ResearcherAnswerDetails();
        Answerdetails = answerdetail.get(j);
        Answer = Answedetail.getAnswer();
        oneLine1.append(researacherName, university, answer);
        oneLine1.append(CSV_SEPARATOR);
        increment j by 1
    write newline;
    bw.newline();
    increment i by 1
```

References

[1] Seyed M Mirtaheri, Mustafa Emre Dinçktürk, Salman Hooshmand, Gregor V Bochmann, Guy-Vincent Jourdan, and Iosif Viorel Onut. Abrief history of web crawlers. *arXiv preprint arXiv:1405.0749*, 2014.

[2] Trupti V Udapure, Ravindra D Kale, and Rajesh C Dharmik. Study of web crawler and its different types. *IOSR Journal of Computer Engineering*, 16(1), 2014.

[3] Siddhartha Reddy. Introduction to web crawling. `http://www.grok.in/blog/2008/06/07/introduction-to-web-crawling/`, Last accessed June 2015.

[4] Classle. Introduction to web crawling 2014. `https://www.classle.net/content-page/introduction-web-crawling`, Last accessed June 2015.

[5] Facebook. `https://www.facebook.com`, Last accessed June 2015.

[6] Linkedin. `http://www.Linkedin.com`, Last accessed May 2015.

[7] Twitter. `https://twitter.com/`, Last accessed April 2015.

[8] Chi-In Wong, Kin-Yeung Wong, Kuong-Wai Ng,Wei Fan, and Kai-Hau Yeung. Design of a crawler for online social networks analysis. *Wseas Transactions on Communications*, 2014.

[9] Shaozhi Ye, Juan Lang, and Felix Wu. Crawling online social graphs. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pages 236–242. IEEE, 2010.

[10] Mike Thelwall and Kayvan Kousha. Academia. edu: Social network or academic network? *Journal of the Association for Information Science and Technology*, 65(4):721–731, 2014.

[11] Reference manager and academic social network for fully-searchable library. `http://www.mendeley.com/`, Last accessed January 2015.

[12] Mark David Slater. Academic knowledge transfer in social networks,(a ph.d. dissertation submitted to the university of california). 2013.

[13] Tejas Desai, Afreen Shariff, Aabid Shariff, Mark Kats, Xiangming Fang, Cynthia Christiano, and Maria Ferris. Tweeting the meeting: an in-depth analysis of twitter activity at kidney week 2011. *PloS one*, 7(7):e40253, 2012.

[14] Douglas RA McKendrick, Grant P Cumming, and Amanda J Lee. Increased use of twitter at a medical conference: A report and a review of the educational opportunities. *Journal of Medical Internet Research*, 14(6), 2012.

[15] MR De Villiers. Academic use of a group on facebook: Initial findings and perceptions. 2010.

[16] TechCrunch By Ingrid Lunden. "who's viewed your posts?" linkedin adds analytics to its publishing platform. `http://techcrunch.com/2015/05/07/whos-viewed-your-posts-linkedin-adds-analytics-to-its-publishing-platform/#.dvmofx:oTGk`, Last accessed May 2015.

[17] Zotero. `https://www.zotero.org/`, Last accessed March 2014.

[18] Citeulike. `http://www.citeulike.org/`, Last accessed March 2014.

[19] Indicators to a certain features on academic social network reserachgate.net. `https://explore.researchgate.net/display/news/2014/08/13/Celebrating+five+million+members+with+free+DOIs`, Last accessed February 2015.

[20] Indicators to a certain features on academic social network researchgate.net. `http://www.alexa.com`, Last accessed March 2014.

[21] Angelika Bullinger, Uta Renken, and Kathrin Moeslein. Understanding online collaboration technology adoption by researchers a model and empirical study. 2011.

[22] Amalia Mas-Bleda, Mike Thelwall, Kayvan Kousha, and Isidro F. Aguillo. Successful researchers publicizing research online: An outlink analysis of european highly cited scientists' personal websites. *Journal of documentation*, 70(1):148–172, 2014.

[23] Kathleen Shearer. A review of emerging models in canadian academic publishing. 2010.

[24] Collections canada. `http://amicus.collectionscanada.gc.ca/thesescanada-bin/Main/BasicSearch?coll=18&l=0&v=1`, Last accessed August 2014.

[25] Theses canada. `http://www.bac-lac.gc.ca/eng/services/theses/Pages/theses-canada.aspx`, Last accessed August 2014.

[26] Proquest. `http://search.proquest.com.ezproxy.lib.ryerson.ca/pqdt/dissertations/fromDatabasesLayer?accountid=13631`, Last accessed September 2014.

[27] Rashmin Babaria, J Saketha Nath, Chiranjib Bhattacharyya, MN Murty, et al. Focused crawling with scalable ordinal regression solvers. In *Proceedings of the 24th international*

conference on Machine learning (A thesis submitted as to Indian Institute of Science-BANGALORE), pages 57–64. ACM, 2007.

[28] Sotiris Batsakis, Euripides GM Petrakis, and Evangelos Milios. Improving the performance of focused web crawlers. *Data & Knowledge Engineering*, 68(10):1001–1013, 2009.

[29] H. Bakshi. Framework for crawling and local event detection using twitter data (doctoral dissertation). 2011.

[30] Zhefeng Xiao, Bo Liu, and Huaping Hu. A facebook crawler based on interaction simulation and mhrw-da. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*, pages 2041–2044. IEEE, 2012.

[31] Masudul Islam, Chen Ding, and Chi-Hung Chi. Personalized recommender system on whom to follow in twitter. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pages 326–333. IEEE, 2014.

[32] SI Mfenyana, N Moroosi, M Thinyane, and SM Scott. Development of a facebook crawler for opinion trend monitoring and analysis purposes: Case study of government service delivery in dwesa. *Development*, 79(17), 2013.

[33] Omar Almousa. Users' classification and usage-pattern identification in academic social networks. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*, pages 1–6. IEEE, 2011.

[34] Arbana Kadriu. Discovering value in academic social networks: A case study in researchgate. In *Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on*, pages 57–62. IEEE, 2013.

[35] Song Q. Chiu D. M. Fu, T. Z. The academic social network.scientometrics. 101(1), 2014.

[36] Mike Thelwall and Kayvan Kousha. Academia. edu: social network or academic network? *Journal of the Association for Information Science and Technology*, 65(4):721–731, 2014.

[37] Encyclopedia britanica, encyclopedia britanica inc. `https://www.britannica.com/`, Last accessed February 2015.

[38] Zi-Lin He, Xue-Song Geng, and Colin Campbell-Hunt. Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a new zealand university. *Research Policy*, 38(2):306–317, 2009.

[39] Omar Almousa. Users' classification and usage-pattern identification in academic social networks. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*, pages 1–6. IEEE, 2011.

[40] Anova table. `https://www.onlinecourses.science.psu.edu/stat414/node/218`, Last accessed December 2014.

[41] Yves Gingras, Vincent Lariviere, Benoît Macaluso, and Jean-Pierre Robitaille. The effects of aging on researchers' publication and citation patterns. *PloS one*, 3(12):e4048, 2008.

[42] Glenn E Hunt, Michelle Cleary, and Garry Walter. Psychiatry and the hirsch h-index: The relationship between journal impact factors and accrued citations. *Harvard review of psychiatry*, 18(4):207–219, 2010.

[43] Dario Taraborelli. Soft peer review: Social software and distributed scientific evaluation. 2008.

[44] Péter Jacsó. A deficiency in the algorithm for calculating the impact factor of scholarly journals: The journal impact factor. *Cortex*, 37(4):590–594, 2001.

[45] Pauline Mattsson, Carl Johan Sundberg, and Patrice Laget. Is correspondence reflected in the author position? a bibliometric study of the relation between corresponding author and byline position. *Scientometrics*, 87(1):99–105, 2011.

[46] & Seeger J. M. Cronenwett, J. L. Criteria for authorship. *Journal of vascular surgery*, 42(4):599, 2005.

[47] Teja Tscharntke, Michael E Hochberg, Tatyana A Rand, Vincent H Resh, and Jochen Krauss. Author sequence and credit for contributions in multiauthored publications. *PLoS biology*, 5(1):e18, 2007.

[48] Serge Galam. Tailor based allocations for multiple authorship: A fractional gh-index. *Scientometrics*, 89(1):365–379, 2011.