University of Southampton

Faculty of Engineering, Science and Mathematics

School of Electronics and Computer Science

Analysis of University Researcher Collaboration Networks Using Co-authorship

by

Jiadi Yao

September 2009

A dissertation submitted in partial fulfilment of the degree of

MSc Web Technology

by examination and dissertation

Abstract

Social network analysis gives evidence for the connections between groups of individuals. It is these connections that channel flow of information and the sharing of knowledge. As universities move towards more interdisciplinary modes of research and funding, an effective network that links its entire cohort of active researchers is vital.

This project conducted a co-authorship network analysis and a path length analysis on a small institutional database. The major advantage of our analysis over other similar work is that we used author's background details in supporting our analysis and generated co-authorship graphs with authors' names and groups.

The network metrics have been compared and contrasted to similar work conducted with large-scale cross-institutional databases in several domains. We found the most of metrics are not affected by the network size and showed that the ECS community is a small-world network with similar knowledge sharing to those communities formed by an entire discipline.

Acknowledgements

I would like to thank my supervisor Dr. Les Carr for all his support and excellent advice throughout the duration of this project. Without his advice and recommendation the completion of this project would not have been possible.

I would also like to thank Dr. Mark Weal for his suggestions and discussion during the project.

Thanks must also go to Ilaria Liccardi for her weekly effort in making sure this work has been on track.

Finally I would like to thank my family for helping me get here in the first place.

Contents

Chapter 1

Introduction

A network consists of nodes and links. Many things can be modelled using a network, for example, power grids, telecommunication networks, the Internet and scientific collaboration. Formulation of a model aims to capture the connectivity and topology of a network, rather than the geometry. A social network is a network with nodes representing people and links representing relationships. Acquaintanceship, co-authorship and collaboration are example relationships that exist between people.

Social network analysis has a history of forty years. It has helped in modelling the speed of infectious disease propagation[23, 29], to understand the importance of certain social relationships in job finding[7] and in realising social relationship between people is a "small world"[10, 25]. However, these analyses do not tell us much about the detailed structure of a network.

Recently, Newman applied network analysis techniques to collaboration networks across domains including physics, biology and computer science[17, 18, 22]. This report attempts to apply the same analytical techniques to a much smaller domain - the School of Electronic and Computer Science(ECS) in the University of Southampton. It aims to compare and contrast the features of the school social network with that of an entire discipline to discover whether the scale of the network leads to differences in network metrics.

The remaining document is organised as follows. Chapter 2 provides a background on network analysis, and similar analysis works are reviewed and evaluated. Chapter 3 conducts the ECS social network analysis, and compares with published results. Chapter 4 focuses on studying the factors that affect the average path length of a network. Chapter 5 discusses the results, and concludes the report. Many large graphs are attached in the appendix.

Chapter 2

Background

The study of social network analysis started back in the 1960s when Stanley Milgram conducted his famous “Six degrees of separation” experiment[25]. Large scale analysis or analysis of large scale social networks did not really begin until recently, as more and more data is stored in a digital form, worldwide networked computers provide ever easier data access, and the processing power of CPU reached a point where even personal computers are able to perform the large scale analysis work within reasonable amount of time.

This chapter reviews the developments of network analysis from paper and pencil experiments in the 1960s, with only hundreds of letters pass through people as probes to digital analysis of networks that consist millions of nodes.

2.1 Previous Network Analysis Work

2.1.1 Small Worlds

The small world phenomenon is the observation that a large network has a small diameter. Although a large network may contain thousands, millions or even billions of nodes, it may be possible to traverse from one side of the network to the other in only a dozen steps. For example, the biggest human constructed network - World Wide Web, contained 8×10^8 documents in 1999, but it was found that on average, one could follow the links on one page, and reach any other in 18 clicks[1]. The acquaintanceship relation between people in the world was also found to have the small world property: the famous phrase “six degrees of separation” [25] means that any two randomly selected people in the world are connected by 6 mutual acquaintances. The network investigated in this report – the co-authorship network – also exhibits the small world property.

Why would we consider the “small world” networks a surprising and interesting phenomenon? Firstly, these are numerically large network - containing billions of nodes. Secondly, there are no central nodes that all other nodes connect to directly, that would allow every node to reach every other though them. Finally, the network is sparse, such that most nodes do not have millions of neighbours to potentially making the network "small". The investigation begins with Milgram's experiment.

Milgram's Experiment The Milgram's work [25] is the earliest empirical study of the human acquaintanceship. The experiment was conducted in the United States in 1969. Letters were given to participating people in Kansas and Nebraska, aimed to be sent to one of two target people in Boston via only those people's acquaintances. The source people were given the basic demographic information about the target person, and told to send the letter only to the people they knew personally, who they think had a better chance of knowing the target person, but themselves should not attempt to get to know the target person. This process was then repeated until the letter either reaches the target or is somehow lost. The successful letters generate chains of recipients. Milgram found that these successful chains had a median length of 6. Later work of Milgram [13] also showed that with sender and receiver in different racial subgroups also give the same result, and hence conclude the entire human society is "small" and suggested the "six degrees of separation".

Milgram's experiment revealed two interesting points. First, the average diameter of the human acquaintance network is only 6; Second, with only *local* information, people are able to find these "short" paths to reach the target person. These sparked some recent research trying to understand these observations.

Reversal Small World Experiment Killworth *et al.* [10], inspired by the Milgram's experiment, conducted their study trying to find patterns of how people choose who to send the letter to next, and potentially discover how people collectively found the short path with only local information. Instead of giving people one target to send the letters to, they gave them 1267 targets, each with the basic location, occupation and ethnic background information, and asked them to write down their choice and the reasons for that first person, among the people they knew, in a potential chain to one of the target. They found that most of the choices were based on the location reason, they also found the choices were mainly friends and acquaintances but not family.

This result appears to explain the underlying structure of the acquaintanceship and how people find short paths. First, people do not choose those "strong" family links when trying to reach someone "far" away. This may be because people know their family members so well that they think they knew and already considered the people their family members may know, so it is meaningless to send it to them. Second, people consider the location to be the most important attributes in reaching someone. This shows that people believe if two people are geographically close, they have a higher chance to know each other. Because Milgram's experiment was a success, then to some degree, this belief must be true.

Watts and Strogatz Model Watts and Strogatz[26] proposed a class of random network to study the small-world phenomenon. Their model starts with a ring of nodes, where each

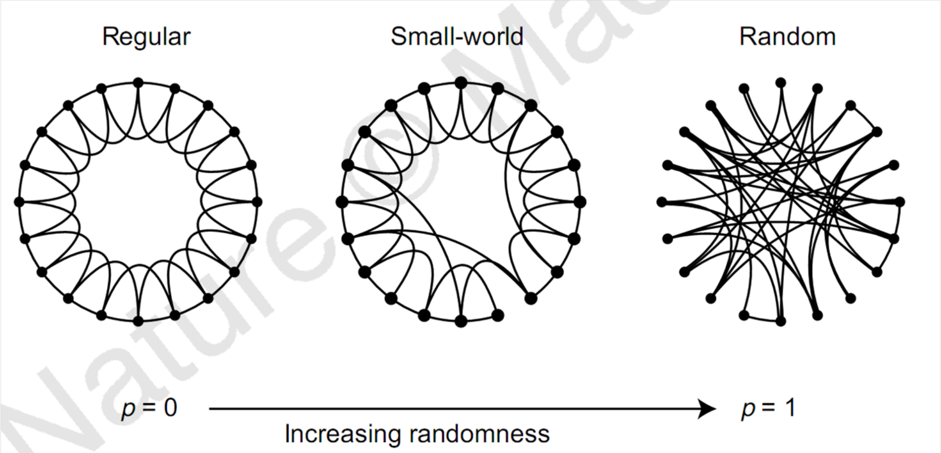


Figure 2.1: Watts and Strogatz Model [26]. Left: the regular ring lattice with no randomness; middle, some randomness introduced when connecting neighbours, the network became "small world"; right, a complete random graph.

node is joined by an edge to their next 2 nearest neighbours, forming a ring lattice (Figure 2.1 left). They call this network "regular" in the sense of no randomness - all the nodes are connected to all their 4 nearest neighbours. They then started to increase the randomness to connect the neighbours, so that some of the local connections became "long" range. The randomness is increased until the resulting network is a complete random graph. They discovered that by introducing a tiny amount of "long" range connection of the regular network it is sufficient to make the network "small". Their model has captured these two crucial parameters of social network: the "regular" models in the acquaintanceship can be thought of people knowing persons close to them in terms of location, and local connections are highly clustered; by introducing some randomness it means occasionally some people know someone far away.

This model allows the randomness of the graph to be controlled, therefore bridges the gap between the pure random graph and the small world graph.

2.1.2 Weak ties

Granovetter[7] claimed that the weak ties are as important as the strong ties in the relationship network. He found the weak ties are the ones that propagate out information to reach a wider range of the network.

He suggested a forbidden triad (figure 2.2), in which if persons A and B, A and C have strong links, it is then unlikely that person B and C do not know each other. For the amount of time A and B, A and C spent together, B and C become less likely to not spend time together. This follows that if the ties within a group are strong, then a single out group tie, which connects to another group must be a weak link. Therefore, in order to reach a wider community, the information must at some stage travel though the weak links. This

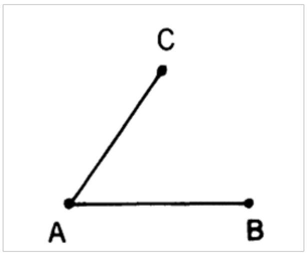


Figure 2.2: The forbidden triad[7]. If AB and AC are strong links, it is unlikely that BC is not connected. This triad should not exist in a social network graph, therefore called "The Forbidden Triad"

weak tie relation was found to help people finding jobs [6, 14, 15], helping firms to create a richer pool of recruits [5], helping to improve cross-functional team effectiveness[24], and it was also demonstrated in Reversal Small World Experiment[10] that people do not choose their family member (strong tie) to reach someone they do not know.

This work is mainly theoretical, but it gave a research focus in the area of social relations. But due to the limitation on the technology - for example, data mainly collected via pen and paper questionnaire, the scale of the network being investigated was limited. Also calculation and simulation without a modern computer was slow and difficult if not impossible. Therefore new research opportunities have emerged as those limitations have been overcome.

2.1.3 Scientific Collaboration Network

Several studies on various scientific networks have been carried out recently. Many of these studies group papers based on the research discipline, for example, Newman [16, 17, 21] studied databases contain papers by mathematics, physics, biomedical and computer science.

The focus of Newman's studies were on the broad statistical properties of the networks, including the number of papers written by authors, the number authors per paper, the numbers of collaborators each author has and so on.

Table 2.1 is a summary of some of the results produced by analysing different subject areas. The biomedical data was from the MEDLINE database, which covers published papers in the field; the physics data is from Los Alamos e-Print Archive and covers primarily theoretical physics and the mathematics data is from The Erdős Number Project.

Papers per author appear to demonstrate similar values across the datasets, but because the mathematics dataset covers about 60 years worth of papers, while the others only about 5 years of papers, so in fact mathematicians produce a lot fewer papers than other areas. The average number of collaborators for the biology dataset is 18.1, while in mathematics it is only around 3.9. Newman claimed that this is caused by the different research modes and groups structure across different fields.

	Biomedical	Physics	Mathematics
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	—
Papers per author	6.4	5.1	6.9
Authors per paper	3.75	2.53	1.45
Average collaborators	18.1	9.7	3.9
Largest component	92%	85%	82%
Average distance	4.6	5.9	7.6
Largest distance	24	20	27

Table 2.1: Some of Newman's analysis results [21]. These results are compared and contrasted in this report. (Number of Paper for Mathematics is not available. It was used Erdős number project as the source data, which counts the author numbers rather than papers)

While these results give an overview of the structure and maybe the patterns and group size of the research work that carried out, there is no focused analysis on, for example, why the average distance in mathematics dataset is 7.6 while the average distance between people in the world is only 6? What exactly caused the distance of a group of mathematicians to be longer?

The main source of the scientific collaboration network was using the co-authorship from the published papers. No study considered the university structure and roles of each individual author. This could be due to those analyses used databases which were cross-institutional, so most authors of the papers do not come from the same university, not mentioning the same school or the same research group. The author's role and group may also not be available to them.

2.1.4 Network Evolution

Another direction of the co-authorship and collaboration network analysis is the evolution of the network.

One of the problems of studying a large-scale evolving network is that there are no clear timestamps on the changes made to the network. For example, the WWW is a constantly changing large-scale network, but the problem is that it is difficult to track the changes and addition of new pages or links at a reasonable rate.

The co-authorship network is both a large scale network which also recorded all changes made to the network. We have already seen the biology collaboration network (Table 2.1) has millions of nodes so it is reasonable large; the time of the changing of the network is defined and recorded clearly by the date of the paper's publication. It is easy to group papers by year in order to observe the evolution of the network across years.

Barabási in their work [2] made some progress in defining a simple model that captures the network's time evolution, although the simulation for various sized datasets show quite different results to the real database results.

Preferential attachment[19] is the main assumption for the forming of the scale-free networks. Most of the scale-free networks are found to be small-world network, such as the WWW, citation network and some social networks. The theory behind a scale-free network is that high degree node has a better chance for those new or low degree nodes to attach to, making the high degree nodes even higher. The majority of nodes in the network only have a low degree.

Jahn[9] considered the meaning of co-authorship, whether to consider the co-authorship to be an author's relational property or as a social event that brings authors together, which would result in two different network modelling - one-mode network and two-mode network respectively. All of the studies discussed previously were based on the one-mode network approach, in which authors represented by nodes, and their relation represented by links. On the other hand, the two-mode network not only models people to people relation, but also the higher level event to event relation - if treating the co-authorship as an event. The two mode network preserves more information from the modelling, but it lacks standard techniques for analysis.

2.2 Relation Data

The previous section introduced and discussed some well known experiments, in this section we will explore various kinds of relation data that can be used to form a network.

2.2.1 Bacon Numbers

This is a famous study of the network of film actors [11, 26] on a database that contains half a million people. Two actors are considered connected if they have been credited with appearance in the same film. The Bacon number is the number of steps each actor needs to reach Kevin Bacon.

Most actors can reach Bacon within 6 steps, but the average path length to reach Bacon is only 2.95, he is not the "best" centre because there are more than 500 people who can produce a smaller average path length for the whole network. The "worst" centre for everyone else to reach produced the average path length 9.

Even though this film actor relation is from people to people, it does not necessarily represent the social structure that underlines the society. So it is unclear whether this particular network has any real social relevance.

2.2.2 World Wide Web

The World Wide Web (WWW) is a source of interesting networks to study. Treating the pages in the WWW as the nodes and hyperlinks in the document as edges, the WWW is the biggest human constructed network in history. To a certain degree, this network may be loosely said to be a social network in the sense that the structure in some way reflects the features of the society that built them, but they do not reflect direct people to people relations. Comparing the metrics between a social network and the WWW yield quite significant differences. To give an example, [1] studied the diameter of the web. Back in 1999, the total number of pages was 800 million, and the mean diameter they found was just above 18. This mean diameter is 3 times larger than the entire human social network - 6 billion nodes with mean diameter 6.

Although the WWW is not the main focus in this study, it has many properties that are unique among various networks. It is a highly dynamic network. New pages become available, old pages die at a high speed, so any snapshot does not describe the network entirely. It is also an explosively expanding network in the current "Internet Age". It is believed to contain 19 billion of pages in 2005 raised to 29 billion in 2007[3]. The recent Web Science Research Initiative[8] is set-up to study the growth of the web.

2.2.3 Co-Citation

As more and more papers are published in a digital form, more content becomes available on-line in digital libraries. Co-citation is therefore heavily used as a source of relations, especially in bibliometrics. A pair of authors is said to be co-cited if they are cited together, regardless of which of their work is cited.

Co-citation offers a powerful method for studying the structure and process of scholarly communication, it has previously used for automatic classification of the relevant history of the field[27], and visualisation of the research front in a specific research field[4].

But the co-citation is a passive relation - two people are linked together because a third person cites them together, it reflects little about those two people's social relation, or even if they know each other at all. Therefore this relation is not considered in this study.

2.2.4 Co-Authorship

In academia, researchers publish papers to contribute to their area of study. Most of these works are carried out collaboratively with a number of researchers, and they publish the work with all of the contributing author's names on. If two researchers names appear on a same paper, they form a *co-authorship* link. Considering researchers as nodes, and co-authorship links as edges, a network can be formed. This is co-authorship network, or research collaboration network.

If two researchers are co-authors, based on working together, they should know each other in person, or may be close friends outside the research activity. This is a common assumption made by other studies [16–18]. But the inverse is not true, close friends have not necessarily written papers together. Therefore, co-authorship does not represent a researcher's entire academic social circle. As a result, co-authorship network cannot simply be treated as a complete social network. Yang *et al.*[28] showed that co-authorship relation is neither a strong relation in interests. Even though, it was considered as the best proxy to a social network[2, 16].

However, research publication is a long process, in the sense that a researcher does not write about their work until the project is finished or towards end, and the paper publication process normally takes months. Therefore, the paper's publication year does not necessarily represent the work undertaken by the researchers during that year.

The biggest advantage of co-authorship compared to others is that it is easier to obtain reasonable amount of data quickly – most of the digital libraries can export these paper's meta-data in batches – therefore leading to a more comprehensive study. While other relationships between academic researchers, for example, 'have been working in the same project', involves manual searching in each funding council's database for current and past projects, which is a slow and intensive work.

The earliest work that uses the co-authorship is in the calculation of the Erdős number. Similar to the Bacon number, Erdős number measures the co-authorship distance to the famous Hungarian mathematician Paul Erdős. Those who published a paper with Erdős have an Erdős number 1, those have published with the co-author of Erdős have an Erdős number 2. But this work only focuses on the micro details – a particular person's distance to Erdős according to his publication, rather than exploring the macro structure of the co-authorship network.

2.3 Summary

This chapter gave a background overview of network analysis from the 1960s until the most recent advances. It reviewed the methodologies and experiments used in those analyses, and identified the limitations. Some key developments were also noted, Milgram's experiments to show the "six degrees of separation", Granovetter recognised the importance of weak ties and Watts Strogatz network model enables the control of network randomness.

A survey was also conducted on the possible relations of creating a social network. Four main relations were considered, co-authorship was the best for representing the ECS collaboration network and the data was very easy to obtain.

Recent co-authorship network analyses were reviewed, some results from those were discussed. These results will be used for comparison later in the report. Even though co-authorship analysis explores some aspects of the relationship, there are vast amounts of other relationships can be acquired and analysed. Recent uprising social network sites, for example, Facebook, Twitter start to change the way people make friends. These could potentially be a more fruitful source of further data.

Chapter 3

Co-Authorship Graph Analysis

This chapter uses the co-authorship relation to study the social network of a local academic community and to investigate whether the network metrics would be significantly different compared to those large databases that cover an entire discipline. Several network graphs are generated by grouping papers on specific criteria; these graphs are analysed, compared and contrasted with the large databases examined by Newman.

Figure A.4 in appendix is a graph showing the intra-group collaboration happened in the ECS. This network contained one thousand nodes, it is tiny compared to the biomedical network contained 1.5 million nodes. But this network is already infeasible to be printed on a reasonable sized paper or to conduct a visual analysis. Figure A.5 in appendix is the inter-group collaboration network within ECS. The length of the links indicates amount of collaborations between two groups. These two figures together provides a visualisation of networks we discuss and compare in this chapter.

Grouping papers together on criteria gives fine control over the important variables, therefore gaining insights and understanding of each aspect's effect on the result. The main criteria are time factor, research group, research topic and individual author.

In order to give a fair comparison with the Newman's data, which includes papers from 1995-1999, two datasets were also included: all the papers between 1995-1999 in the EPrints and all papers by one of the groups in 1995 - 1999.

In the rest of the report, the word node and author are used interchangeably; author's collaborator and node degree are used interchangeably.

7 properties of the graph are calculated and compared. These include:

- The number of papers - Total papers in the dataset, measures the size of the data source.
- The number of authors - Total unique authors from the papers, the number of nodes in the network.

- The average path length - The average of the shortest distance between every connected pair of nodes. Measures the mean separation of the nodes in the network. Unconnected pairs are excluded.
- The diameter of the network - The minimum distance between all pairs of nodes. Measures how far apart are the most distant pair.
- The author's average collaboration - The average number of all author's collaborators; in the network sense, the average of all nodes' degree. Measures how collaborative the authors are.
- The average number of authors per paper - measures the average size of the collaboration.
- The size of the largest component - The percentage of nodes connected to the largest component. Measures how connected the network is.

3.1 Data Collection and Methods

There are two EPrints systems in University of Southampton, ECS-wide and University wide. Both systems store thousands of papers and have new deposits almost every day. This study used papers from the ECS EPrints repository.

The paper upload for the ECS repository is mainly done by individual researchers. Uncontrollable errors may present in the data. Some of these errors are obvious, for example, there is one paper published in year 0008; errors related to the author's identity are more difficult to find and correct. During the analysis, we found that some authors do not have ID in some papers they publish, so they have not been counted even though they should. The way this error could affect the result is making the network slightly smaller. Using the assigned ID to identify each author is much more accurate compared to using author's name. In Newman's research [21] the difference between the number of authors identified using full name and first initial can be as large as half a million, introducing a major source of error.

The EPrints repository system is a fairly recent technology, the first version of the software was developed in 2000. Before ECS start using EPrints system, all publications' metadata were stored in a database and maintained by a single person. By browsing the paper by year, we notice that papers before 1990 are significantly fewer per year and get fewer as it goes further back. But in the analysis, all papers are used regardless what year they were published because the author's relationship and the larger amount of data is more important.

Author's Group, Role The ECS website has a staff list page with their roles and groups they work in. This paper is used as the data source by the program to look up author's group and role.

	Papers	Authors	APL	Diameter	Collaborators	Authors per paper	Largest Component
ECS-All (1965-2009)	13,933	1,028	4.05	10	6.67	1.82	95%
Year 1995-1999	3591	200	6.18	18	2.68	1.37	64%
Year 2008	941	424	5.45	13	4.52	2.45	81%
IAM-All (1972-2009)	2,936	315	2.73	5	8.73	2.20	99%
IAM (1995-1999)	562	55	2.37	5	3.78	1.77	80%
Research Topic (1994-2009)	1,030	287	2.93	7	7.69	2.74	90%
Researcher (1993-2009)	76	47	1.78	2	9.53	4.32	100%
Biomedical (1995-1999)	2,163,923	1,520,251	4.6	24	18.1	3.754	92.6%
High Energy Physics (1995-1999)	66,652	56,627	4.0	19	173	8.96	88.7%
Computer Science (1995-1999)	13,169	11,994	9.7	31	3.59	2.22	57.2%
Mathematics (1940-1999)	—	253,339	7.57	27	3.92	1.45	82%

Table 3.1: Top half: The results of analysing 7 specially chosen datasets; bottom half: results by Newman[16] on the domain wide databases. **APL**: Average path length. The average of the shortest distance between every connected pair of nodes. Measures the mean separation of the nodes in the network. Unconnected pairs are excluded. **Diameter**: The longest distance between all pairs of nodes. Measures how far apart for the most distant pair. **Collaborators**: The author's average collaborator. Measures how collaborative the authors are. **Authors per paper**: measures the average size of the collaboration. **Largest component**: The percentage of the nodes connects to the largest component. Measures how connected the network is.

3.1.1 Programming Consideration

In this project, there is no software product produced at the end. The programming language is primarily used as a tool for data extraction, calculation and graph drawing. Even though, it is important to use a programming language that allows fast develop - test cycles and supports external libraries. Perl was chosen because it is an interpreted language for fast development and testing. It supports a wealth of external libraries and supports graphviz, which is a useful graph drawing tool. But the downside is that the author does not have previous experience of using Perl, there is a risk of not be able to learn the language quickly enough for using in this project. A back up language - C - is also considered. The author has extensive experience in C, and used it to developed several sizeable projects, the problem with is that it is a compiled language, it may runs faster, but the strict syntax and lack of an easy-to-use external library may cause trouble.

Scripts There is one main script used in this study, it is about 800 lines long. It takes EPrint3 XML data file and performs calculation and produces graphs based on the network described in that data file. Performing different functions and various calculations is by ad-hoc change of the script, but they can be easily developed into a fully automatic network analysis program. The running time of the script is mostly dependent on the input network size and the work needs to be performed. To calculate some simple property on a small network only takes a fraction of a second on a 1.8Ghz computer, but to calculate APL or generate graphs on a large network can take minutes.

3.2 All Papers in ECS EPrints to Date (1965-2009)

This dataset(ECS-All) includes all papers in the ECS EPrints repository. It contains 13933 papers since 1965. Analysis on this dataset provides a global view of collaborations between ECS members in these years. (Table 3.1)

Comparing the result with Newman's co-authorship analysis [16], 13933 papers by 1028 authors is significantly more papers per head in ECS than the computer science database, which published 13169 papers by 11994 authors. This indicates the researchers in ECS take more initiative in publishing their work. But this comparison may contain errors introduced when counting the authors. We used unique IDs in identifying authors, and the majority of the authors had a valid ID, however some authors were dropped out. Newman's analysis on the other hand, used author's name as the identifier. Due to the ambiguity in the printing of the authors' name, one author can be recognised as several and several can be confused into one, lead to errors in estimating total authors.

The average path length (APL) of ECS-ALL dataset is 4.05. It is difficult to make any conclusion on the APL with only these overview data as there are conflicting observations. Although the network size of ECS-ALL is the smallest – with only 1028 nodes, the high energy physics, which has 50 times more nodes in the network, produced a smaller APL –

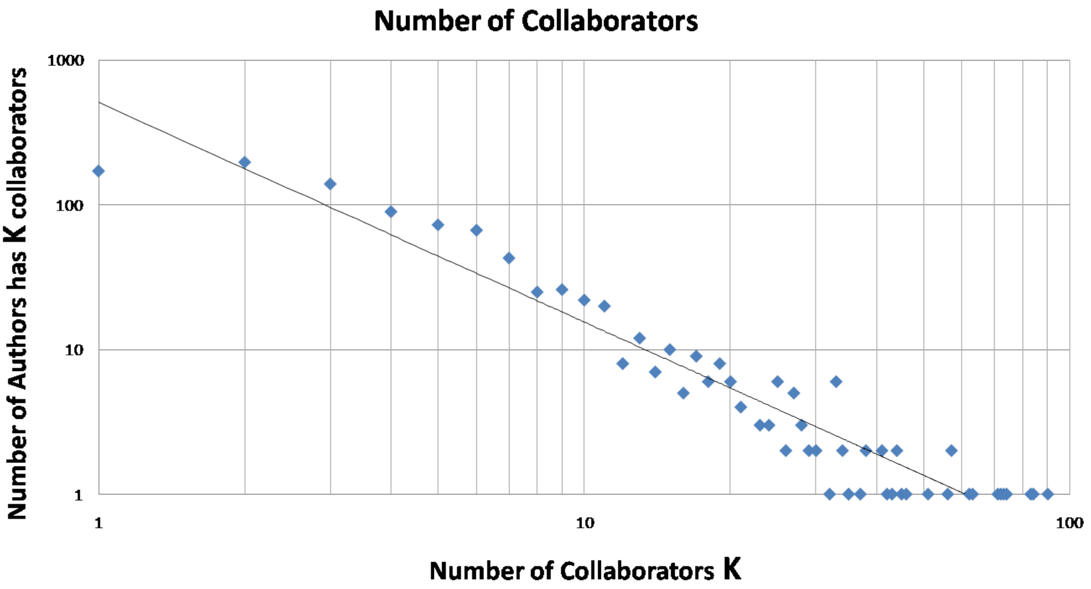


Figure 3.1: Number of collaborators plotted on the logarithmic scale, the straight line is the best fit. The degree distribution in this smaller domain exhibits the power law.

4. If we compare this to computer science dataset, its APL is 9.7 – more than double that of ECS-ALL – while its network size is 10 times larger. Two larger networks produced one larger APL and one smaller APL. There are studies [12, 20] which show that the rate of APL increase in a small world network is slower than the logarithmic of the nodes in the network. We cannot see this pattern here. In chapter 4, we have examined the data and trying to find the factors affecting the APL.

The average author's collaborators is 6.67, almost double the collaborator's in computer science and mathematics. But it is significantly lower than high energy physics and biomedical. This value reflects the size of the project in the specific domain. High energy physics sometimes requires hundreds of people to complete one project. The higher collaboration size in ECS compared with computer science or mathematics is caused by larger practical electronic projects.

Figure 3.1 shows the number of collaborators on a logarithmic scale. It exhibits a power law distribution - the majority of authors have just few collaborators and the majority of collaborators goes to a few people. This result is in line with Newman's work [17].

3.3 Papers by Year

Two periods are chosen for analysis, 1995-1999 and 2008 (Table 3.1). The formal set allows a fair comparison with Newman's dataset of the same period. The 2008 dataset is the latest complete year of ECS's publication, and provides an analysis on the most recent publication pattern in this school.

Although ECS published only 3591 papers during 1995-1999, this is far fewer than was recorded in the computer science database. The number of authors who produced these papers were also a lot less. On average, each author had published nearly 18 papers during the 5 year period[1] while the authors in computer science database only published just above 1. The research output rate of the scientists in ECS is much higher than the authors in the computer science database.

The higher rate of research output may also relate to short APL. Between 1995-1999, the APL of the researchers in ECS was 6.18, over 3 steps shorter than the computer science database. This makes the communication and information flow through the network faster, resulting in a more productive and efficient community here in ECS than the computer science community Newman studied.

3.4 Papers by Group

The Intelligence, Agents, Multimedia (IAM) Group contains the most researchers in the ECS. This group had contributed to over a fifth of papers among the total 22 research groups (Table 3.1). Although this group overall published the most papers, it certainly is not always have been. The IAM-All dataset includes papers across the 37 years, and on average authors only published 1 paper ever 4 years, far less than the ECS-All average. This could due to the incomplete data before the year 2000.

The APL in the IAM-All dataset is much shorter than others. Its average path length is 2.73, only half of the length in 1995-1999 dataset. But the number of authors in IAM is in fact quite a lot more than in the 1005-1999 dataset. The diameter is also much shorter, less than a third of the 1995-1999 dataset. The average number of collaborators within this group reached 8.73, higher than 6.67, the overall number of collaborators, and several fold higher than in year 1995-1999. Almost all authors within IAM-All are connected, the largest component reached 99%. All those metrics indicates that people have been working closely together within the group.

Those particular high collaborative authors, for example Professor David De Roure, who had 90 co-authors across the school, has 76 of those co-authors come from the IAM-All dataset alone. Further analysis showed that 51 of the 90 co-authors belong to the IAM. More than half of the collaborators came from the same group confirming that the collaborations within the group are far stronger than across group.

The short APL, short diameter and the high number of collaborators as well as almost all-connected collaboration network shows that researchers are working closely together

[1] The research output or paper per author used here is only an estimation, using the total number of paper divide by total number of authors. Authors publish papers collaboratively, so one paper can contain several authors, therefore the actual paper per author would be larger

within the group. As a result of all these, one would expect IAM's publication rate should more than the average, but it was not as already shown above.

We discuss the effect of the path length in the research community in chapter 4.

3.5 Papers by Research Topic

The research topic dataset is generated by using the EPrints query engine with the keywords "Semantic Web". More than 1000 papers returned indicating this is a heavily researched topic in this school(Table 3.1). The APL was 2.93, which is in the similar range of a group-based dataset. The diameter of the collaboration network is 7, only slightly greater than IAM group. The largest connected component also reaches 90%. All these properties indicate the similarity between a research-topic and a research-group. This in fact reflect the reality. Groups are set up based on the similar research direction, so a paper published by the same group should mostly be in the same field; while the search on a specific research-topic should bring back all the papers published on that topic. Similarity in the author collaboration graph would be expected.

But the main flaw in ECS EPrints repository is the imperfect query engine and "Semantic Web" is only a phrase. There can be many papers returned nothing to do with "Semantic Web".

3.6 Papers Containing a Specific Researcher

A network was generated for author Dr Mark Weal. The APL in this network is only 1.78 and the diameter is just 2. Clearly, all the authors should connect directly to Mark to make the diameter 2. There should be a significant number of pairs of authors who have direct connections between themselves in order to bring the APL within 2. The entire network is connected, so the largest connected component contains 100% of authors.

The network generated by papers that containing a certain researcher is in general small - only a dozen nodes. Therefore, we can analyse them by looking at the network directly (Figure A.1). This graph is highly centralised, we are able to observe that all of the authors are connected to Mark Weal. No author has been left out of the main component. Additionally, the graph gives a direct sense of how collaborative this researcher is and his minimum social circle within the school. It also shows the research groups the author had collaborated with, and hence the expertise needed by the researcher to carry out the work. In this case Dr Mark Weal had worked most closely with the researchers from the IAM group as well as the Learning Society Lab(LSL). He at least knew 46 people quite well in the school.

In contrast, Ian Heath (Figure 3.2) had a much narrower collaboration circle in the school.

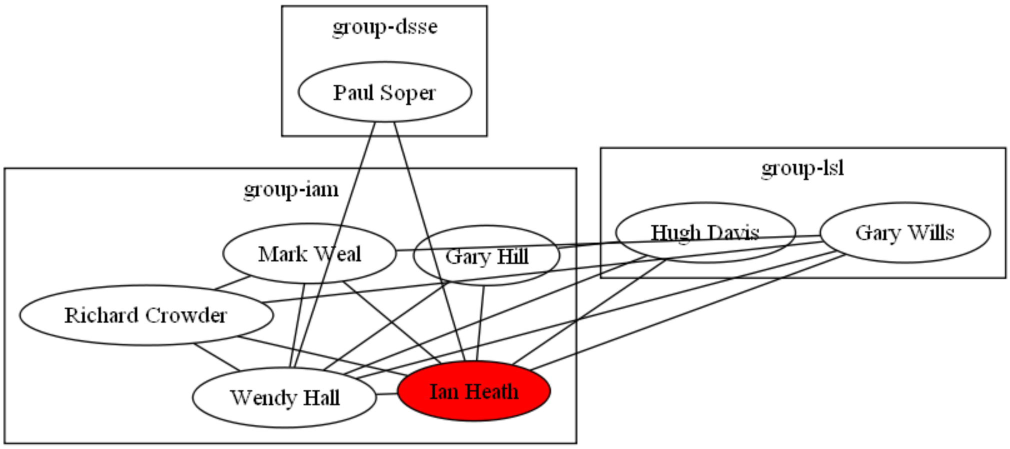


Figure 3.2: Ian Heath's collaboration network. Comparing this network with Mark Weal's (appendix A.1) we immediately see how small Ian's collaboration is. The collaboration network size may be due to the different research work each author was involved in.

3.7 Summary

This chapter analysed, compared and contrasted several network graphs generated by various criteria with the published results.

Compared to the computer science, biomedical and high-energy physics databases, ECS produces significantly more papers per author. Although some errors may have been introduced from over-estimating the number of authors in those databases due to ambiguous in identifying authors by their names. The network metrics calculated with this much smaller scale database were not significantly different from those cross-institute domain-wide databases.

The APL of a network is not directly correlated to the number of nodes. One larger network - high-energy physics - produced a smaller APL, while other larger networks produced longer APLs than the ECS-All dataset. So there must be factors other than size of the network that affects the APL.

The ECS's collaboration network follows the power law distribution. Hundreds of authors have only one or two collaborators, while only several authors have a large number of collaborators. This observation is in line with Newman's results [17].

The group-based dataset generated a network that produced a short APL, a small diameter, a high average degree and an almost all-connected network, but the publication rate from this group just below average. This raises a question of what factors makes a group more productive? The APL and the relevant network property are studied in more depth in section 4.

Papers grouped by research topic produced a network similar to the group-based dataset - the APL, diameter, collaborator, and the largest connected component gave similar values to the IAM-All dataset. This verified that the IAM group is working and producing papers in a certain area.

Finally, papers containing a certain researcher generated a centralised network, where every author connects to that researcher directly. These networks are small in terms of the number of nodes, therefore we can look at the network graph directly. By visual analysis of the nodes and the connection of edges, we were able to see the author's collaboration circle size.

Chapter 4

Path Length Analysis

Chapter 3 gave an overview of how the co-authorship graph appear when grouping papers in certain criteria, highlighting the average path length of a network. As already seen, a collaboration network with 50 times more nodes produced a shorter APL than the ECS-All dataset; while a 1500 times larger network produced a APL 0.5 shorter APL than the ECS-All dataset. Clearly, the length of the APL is not strongly related to the size of the network.

In this chapter the focus is on determining the factors that affect the length of a network's APL.

4.1 Distribution of Path Length

To understand why the APL is a certain length, it is necessary to investigate the length distribution. Figure 4.1 shows the path length distribution percentage of each dataset.

The 2008 dataset (dark blue) distributes evenly over a longer range compared to the other datasets. More than 80% of its pairs have lengths between 3 and 8. The highest peak length is 4, but it has only 1% more pairs than the second highest. This dataset also has the longest path length – 13 – across the datasets.

The IAM dataset(red) produced a more concentrated distribution, with nearly 60% of the pairs of length 3. This alone would make the average not far from 3. The longest path is 5 steps, it is too little to be shown on that figure. The distribution shape of this dataset looks like normal distribution – high in the middle, low at both ends.

The topic based dataset(green) gives a similar distribution as group based dataset, the shapes look like a normal distribution, where most pairs have length 3. The longest length is slightly longer than IAM dataset – 7, but 97% of the authors are connected with 4 steps. The average path length would certainly be below 4.

The researcher-based dataset(purple) gives an expected short path-length distribution. The

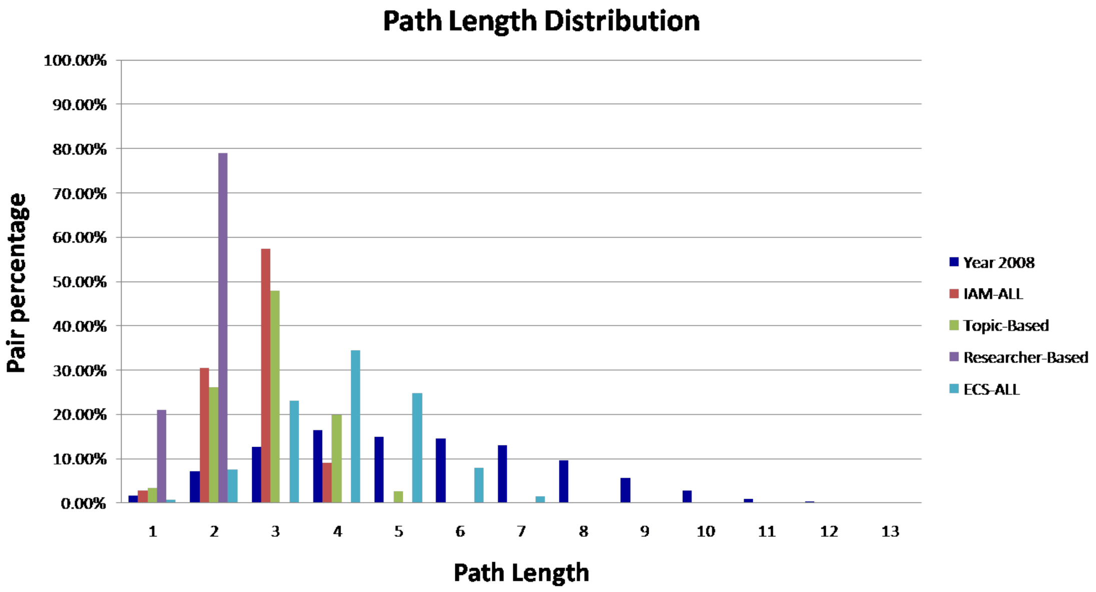


Figure 4.1: Path length distribution percentage of each dataset. The pair percentage distributes evenly over a range of path lengths for a network had a large APL, e.g., Year 2008. For those short APL datasets, the distribution is more biased towards a particular length. E.g. IAM-All dataset had nearly 60% of pairs at length 3; Researcher-based dataset had almost 80% of pairs at length 2.

longest path is only 2 steps, as everyone can reach everyone else via the centre researcher. The length 1 pairs reached 20%, which would bring the average path length within 2.

The ECS-All dataset(light blue) produced the most balanced shape: peaks at length 4 with over 30% of pairs, then gradually the pair number lowers as the path length gets larger and smaller. This balanced shape may result by its larger dataset. From this distribution shape, it is certain the average would lay around 4.

From the above results, it appears that for a network has a large APL, the pair percentage distributes evenly across a range of lengths. There was no single length that had more than 20% of the pairs. But for those short APL datasets, the distribution is more biased towards a particular length. In the following sections, we explore the reasons for a length of pairs in a network and how a node's degree affect the path length.

4.2 Path Length and Node Degree

Figure 4.2 is the average author's collaborators against their positions in the chain. The 13-step pairs were chosen because they are the longest paths in all the datasets. From this figure, we can see that the authors' average collaborator number at the ends is significantly lower than authors towards the centre. The author at position 1 and 14 only have 2, while authors at position 4,7,8 and 11 have 20 or more. As a result, authors who do not have

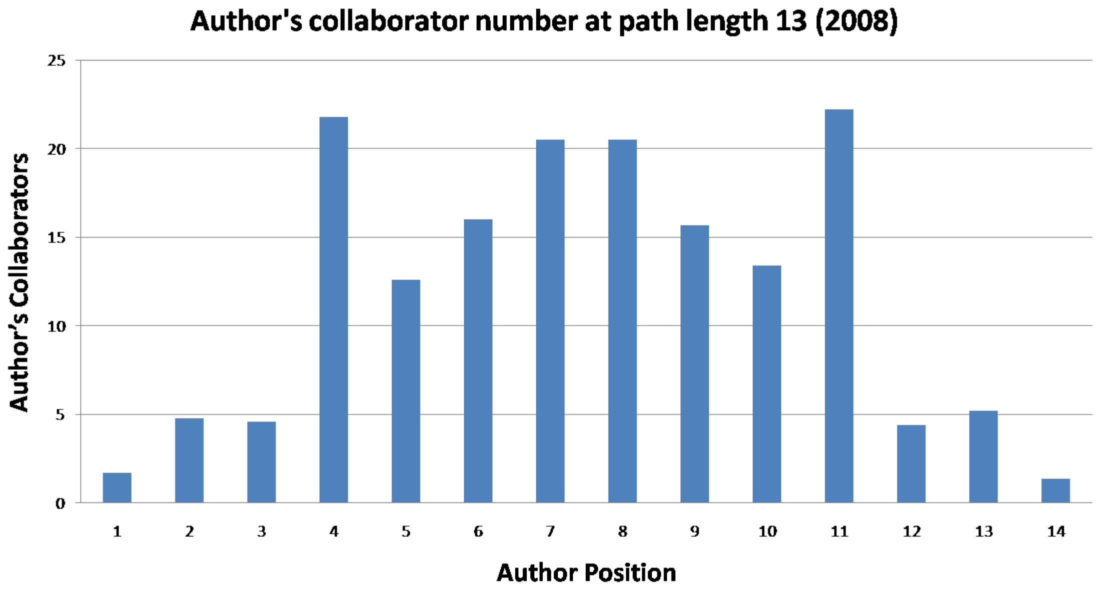


Figure 4.2: The average of 30 pairs of 13 step length author's collaborators against their positions in the chain. The number of collaborators for the authors at each end is a lot fewer than for the authors towards centre.

many collaborators have extended the length by around 2. This is investigated in detail in section 4.3.

We also notice that the author's collaborator number do not always increase as their position move towards the centre, which is different from the path length distribution figure (Figure 4.1). The authors have most collaborators appears towards the centre, but not necessarily in the centre of a path.

In this school-wide repository, and probably also true in larger repositories, we found that nearly 70% of the authors who only collaborated with one or two other authors are post-graduates, short-stay and visiting researchers. Post-graduates are just entering the field and starting to publish papers with one or two of their supervisors. Short-stayed researchers or visiting academic staffs have not been in the ECS long enough to collaborate with many people. (Some error may be present due to the incomplete data, as nearly two fifths of the authors who have one or two collaborators do not have a role attached.)

The node degree is much related to the path length, especially those end nodes. It appears that those degree 1 authors contributed to at least length 2 to the APL, which we explore next.

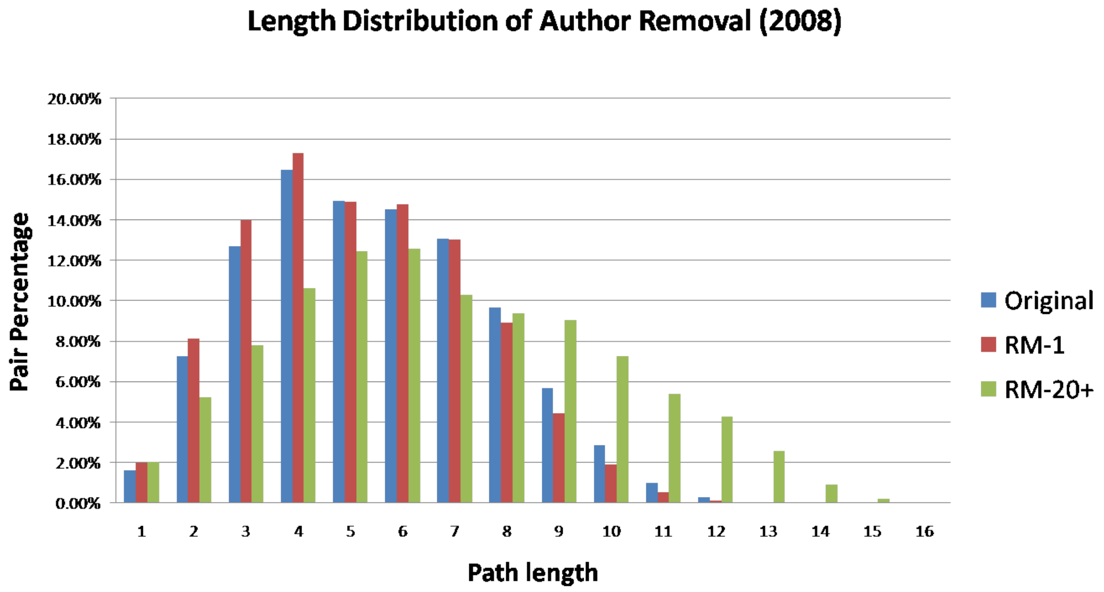


Figure 4.3: The path length distribution of the original 2008 dataset (Original), the 2008 dataset after removing degree 1 nodes (RM-1) and the 2008 dataset after removing node with degree greater than 20(RM-20+). Removing degree 1 nodes only slightly shortens the path length, but removing node with degree more than 20 significantly extended the path length.

4.3 Path Length affected by Node Removal

This section experimented on removing certain nodes from a dataset, and trying to find the impact on the APL.

Figure 4.3 shows the path length distribution of the 2008 dataset after removing degree 1 nodes and nodes with degree greater than 20. The pair percentage in RM-1(Dataset that removed degree 1 nodes, red bars) is mostly higher than the original(blue bars) before length 6, and then lower than the original after 7. Although the longest path is still 13, which is not shortened by the removal, the percentage of pairs longer than 11 has reduced to negligible amount. Overall, the amount of short lengths was increased, while the long lengths was reduced. This would certainly lead to a shorter APL – 5.21. Unfortunately, this new APL only shortened a fraction of the expected length.

The distribution of RM-20+ (Dataset that removed degree more than 20, green bars in figure 4.3) spread out more than the original. All the pair percentage before length 8 has reduced, some short path lengths, for example, 3 and 4 have reduced nearly half of the original percentage. The pair percentage has increased dramatically after length 8 and the longest path length now becomes 16 from the original 13. From this distribution figure alone, we would expect a significant path length increase. The APL for RM-20+ dataset is 6.79.

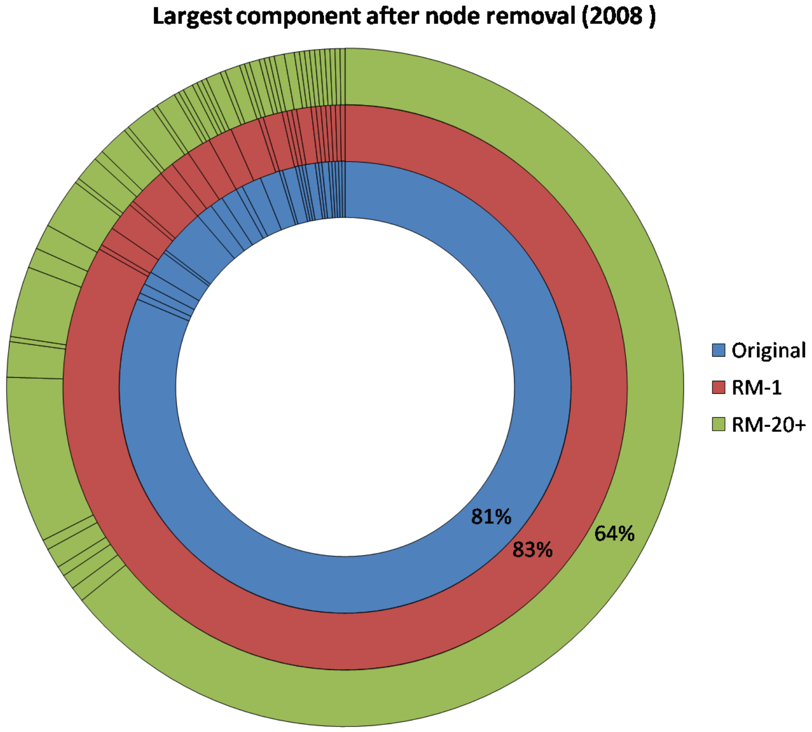


Figure 4.4: The connected component graph of three datasets: the 2008 dataset(Original), the 2008 dataset removed degree 1 nodes (RM-1) and the 2008 dataset removed nodes with degree more than 20 (RM-20+). Each ring represents a different dataset. The sections in the ring represent the connected components in the network. The largest connected component in each network is the largest section in the ring. Removing degree 1 nodes only slightly increase the size of the largest connected component, from 81% to 83%, but removing nodes with degree more than 20 significantly fragmented the connected components as well as decreased the size of the largest connected component. The largest connected component in the RM-20+ ring reduced to only containing 64% of the nodes in the network, while a lot of nodes become disconnected and forming many smaller components.

Figure 4.4 shows the connected component graph of the three datasets – the 2008 dataset, 2008 dataset with node 1 removed, and with node more than 20 removed. Removing the 8 high degree nodes fragmented the largest connected components. The largest component has reduced from containing 81% of the nodes of the original dataset(blue ring) down to only containing 64% of the RM-20+ dataset(green ring), the total number of components has risen from 27 to 48. But removing the low degree nodes has not made such an impact. The largest connected component of the RM-1 dataset (red ring) only 2% larger than the original, and the number of component is 3 less.

In a group-based dataset, the effect of removing high degree node is magnified. Figure 4.5 is the path length distribution of IAM-All with degree 1 and degree 20 and larger nodes removed. From the figure, the contrast between the original dataset(blue) and the dataset after removing the nodes with degree more than 20 (green) is so large. The RM-20+ dataset have expanded to length 17 from the original 4. The percentage of the pairs in

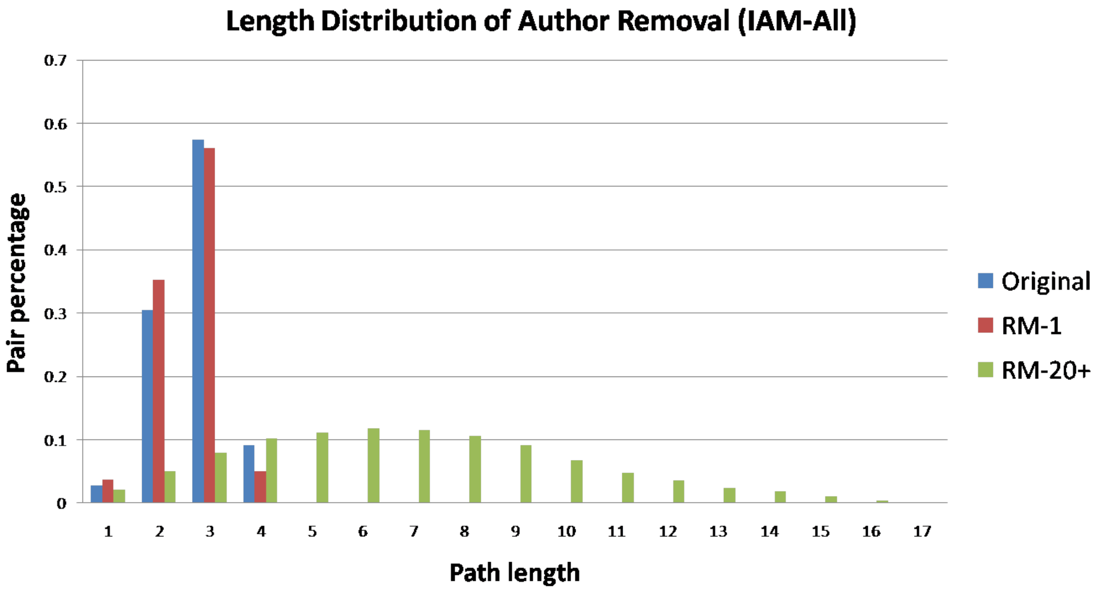


Figure 4.5: The path length distribution of the original IAM-All dataset (Original), the IAM-All dataset after removing degree 1 nodes (RM-1) and the IAM-All dataset after removing node with degree greater than 20(RM-20+). A magnified effect compared to figure 4.3. Removing nodes with degree more than 20 increased the longest path length by 4 times.

length 2 and 3 has shrunk to only a fraction of the original. The APL of the RM-20+ dataset is calculated to be 6.90, thus increased 4.17 from the original. While removing the high degree nodes made such a large difference, the effect of removing degree 1 nodes was only minor. The pair percentage slightly increased in length 1 and 2, and slightly reduced in length 3 and 4, making the APL only 0.11 shorter than the original. Removing the high degree node significantly fragmented the connected component in the IAM dataset. The largest component size has dropped from containing 99% of authors down to 69%. The number of connected component rose 25 times from 3 to 75.

4.3.1 Node Removal Conclusions

This section showed the important role of the high degree nodes in pulling together the network. Removing them caused a signification impact both in terms of the extended path length and the fragmentation of the connected component. The high degree nodes appear to be much more important in a group-based dataset, removing them almost devastated the entire group. As a result, high degree authors become the weak points in the network.

On the other hand, those low degree nodes played a much less important role in the APL length. Removing 8 times more low degree nodes than the high degree nodes only made the graph a fraction shorter.

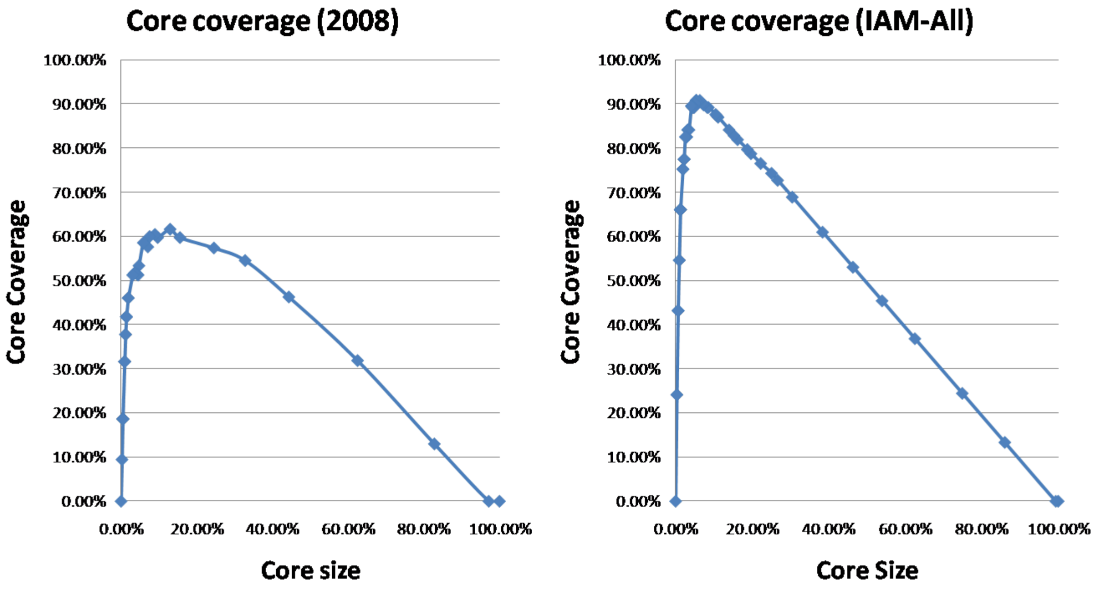


Figure 4.6: Left: Core size against the core coverage in the 2008 dataset. Right: Core size against the core coverage in the IAM-All dataset. IAM-All is a better connected graph as it has a smaller core size than 2008 dataset while covered larger percentage of nodes.

In the next section, the focus is on defining the "high degree" nodes. Given a network, above what degree should be called "high"? We will call the set of "high degree" nodes the *Core* of the network.

4.4 Finding the Core of the ECS Network

The "high degree" in the previous section was not clearly defined, it assumed the nodes with degrees above 20 were high degree ones. The main characteristics of a high degree node is it connect directly to many other nodes. A core is the set of nodes that connects directly to most of the other nodes.

To make the expression easier, the set of nodes that is connected directly to the core nodes are called *core coverage*. If a node is considered to be core, it is no longer counted towards coverage.

The control of the core size is thresholding the node degree. The smallest core is the set of highest degree nodes in a dataset. The threshold is then lowered to include more nodes in the core.

Figure 4.6 shows the core size against the core coverage using both the 2008 and the IAM-All dataset. In both figures, the core coverage increase rapidly when the core is small and only contains a few high degree nodes. The 2008 dataset has a peak at just above 60% of the coverage, with 13% of the core size, while the IAM-All dataset peaks at 90% with a smaller 7% core size.

4.4.1 Conclusion on Finding the ECS Core

The IAM-All dataset had a smaller core while covering a wider part of the network indicating the community formed by it is more *centralised* than the community formed by 2008 dataset. As a result, removing the same high degree nodes from the IAM-All dataset separated the network more when compared to the 2008 dataset. A centralised network may have motivated the communication, and improved knowledge sharing. The fact that every piece of information have to go through the central nodes creates a bottleneck as well as a "point of failure" in the network. This reveals the potential weakness in the structure of group-based networks in the ECS.

4.5 Evolving Path Length over the Time

In previous sections, we have analysed the relationship between the path length and node degree. We had strong evidence of the high degree nodes in making network small. In this section, we change the angle of the analysis. We look at the time factors that may cause the change of the path length.

There were some researches considering the network evolution over time. Barabási et. al.[2] used co-authorship data over the years to capture the node and edge addition to the network, and developed a network model. Newman [19] studied the probability of future collaboration between researchers and concluded that the probability of collaboration is strongly positively correlated with their number of previous collaborations, and their number of previous collaborators. But these studies did not consider the path length or "distance" change between researchers over the time, which we investigate in this section.

In the rest of the report, we use "distance" between authors to mean the number of steps these two authors need to go though their connected collaborators to reach each other.

Long-Path Change when Extending Time There are 30 pairs of authors who have path length 13 in the 2008 dataset. One pair is randomly picked to investigate in detail. This pair of authors is used as the end points and queried for in other year based datasets to find out their distance in each. The dataset used are 2007-2008, 2006-2008, 2005-2008 and ECS-All. In figure 4.7 (left), the distance between this pair of authors decreased steadily as the dataset include more and more years until the shortest distance between them – 5 is reached. Figure 4.7 right shows the paper and author number for these datasets. Both numbers increase as the dataset gets larger. There is a slight jump from 05-08 to ECS-All in the paper number but no sudden increase in author number.

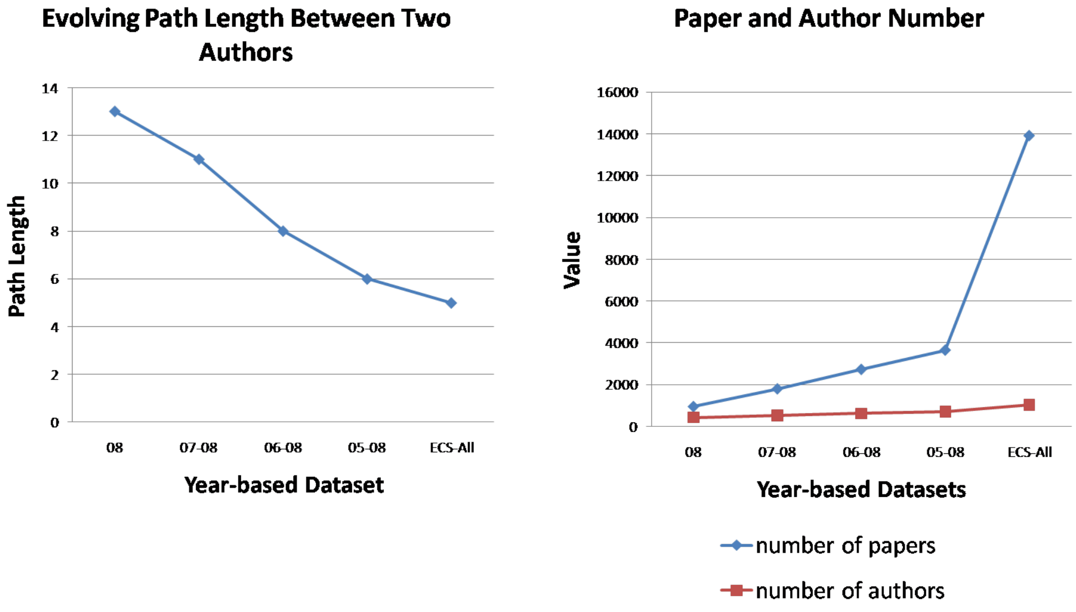


Figure 4.7: Left: The path length for a fixed pair of author decrease as the dataset gets larger. Right: The paper and author size as the dataset gets larger.

Path Length from Research Group Level Looking at the research groups of these authors who belong in the ECS, these paths split into connected research groups – links go though several authors in the same group, then move on to next group until reaching the target. As the dataset increases in years, the group members "expand" to either omitting the neighbours and link directly to next one on the chain (See Appendix A.2. e.g. Chris Harris omitted Bob Damper and Steve Gunn and connected direct to John shawe-Taylor in 2005); or a new author may replace a number of old authors in the same group (e.g. Trung Huynh replaced Nigel Shadbolt and Wendy Hall in 2007).

The number of research groups involved in the path has dropped from 5 in year 2008 down to 2 in ECS-All. This means that these two groups – ISIS group and EPE group – have collaborated in the past, but not every year. The 2008 group-collaboration network graph (Appendix A.6) reveals two interesting points. Firstly, this is a weakly connected graph, only 39 out of total possible 231 edges are connected. The most connected group is group-ESD, but only directly connects to 11 groups out of total 22 valid groups. The fact that even at the group level, not every group has collaborated with any other, would mean a high APL at author level. The APL for the group level graph is 2.32. How this macro value is related to the micro author level APL value may be an interesting further study.

Secondly, we notice the group-EPE and group-ISIS, which had 13-step path, are actually connected in 2008. The question is then why it had to go though other groups in order to connect those two authors, rather than utilise this direct link?

Figure A.3 in appendix shows the collaboration at author level between two groups in 2008.

The two red nodes are the authors we want to connect. From the graph, we can see that the ISIS group has split into two big parts(marked out by the red dotted line), however only one part is connected directly with the EPE group.

Short-Path Change When Limiting by Time From the ECS-ALL dataset, the path length 4 was chosen because it has the most pairs. These pairs are queried against the 2008 dataset to find out their new path length with this single-year dataset. There in total 163,291 pairs of authors who have path length 4 in ECS-ALL, but most of these pairs were no longer connected because either one of the author or both of the authors do not exist. There were only 18,535 pairs (11%) still connected, and their APL is 5.93, increased from 4.

4.5.1 Evolving Path Length Conclusion

The path length of a *fixed pair* is directly proportional to the size of the network. More nodes in a network allows a specific node to reach another in a shorter distance than a smaller network. But as we have seen in chapter 3 that the APL of a larger network is not necessarily shorter than a smaller network. Therefore, the distance between the additional pairs in a larger network must have offset the APL reduction.

There are large groups in ECS are split into smaller communities, this separation made a otherwise short path into the longest path. So, be able to locate and recognise these obstruction would promote a better connected scientific community.

4.6 Summary

This chapter attempted to identify the factors that affected a network's APL.

We have seen that for a network which has a large APL, the pair percentage distributes evenly over a range of path lengths, while no single path length had over 20% of the pairs. But for those short APL datasets, the distribution is more biased towards a particular length. For example, IAM-All dataset had nearly 60% of pairs at length 3; Researcher-based dataset had almost 80% of pairs at length 2.

We then investigated the author's average number of collaborators in regarding their positions in a chain and we found that those less collaborative authors tend to appear at the ends of a path, while those highly collaborative authors only appear towards the centre. In this particular repository, we found the authors that only have 1 or 2 collaborators are mostly post-graduate student, short-stay and visiting researchers. As a result, we believed that those low collaborative authors extended the APL of a network for about 2. However, further analysis showed that the low degree nodes in a network do not affect the APL as much as the high degree nodes do. Removing 60 degree 1 nodes from a 400 nodes network did not shorten the APL nearly as much as the removal of 8 top degree nodes which had extended the APL of the network. The impact of removing high degree nodes is stronger in a group-based dataset than in a year-based dataset. Therefore we also investigated the portion of network those high degree nodes can reach in 1 step. The results show a smaller core in the group-based dataset and a wider coverage, while in year-based dataset, a larger core was needed but covered only a narrower set of nodes.

Finally, we analysed the relationship between the path length and the size of dataset varying by time. The experiments show that as the dataset to include more papers, the distance between a pair of far away authors would reduce. The inverse is also true: in a dataset which contains multiple years of paper, when limiting the dataset size by year, those originally short-distance pairs became far away. If we view each path from the research group level, we found that the path goes though group after group. We also found that many groups do not form one connected component, making the path length between some pair of authors very long even the groups the authors belong to had direct collaboration.

Chapter 5

Discussion and Conclusion

This report applied analytical techniques to papers in the ECS EPrints repository to compare and contrast the features of the school social network with that of an entire discipline to find out whether the scale of the network would change the network metrics. It further analysed the path length in the network in the attempt to find the factors that affect the length of a network's average path length(APL).

The average path length measures the distance between nodes in the network. In a small world network like the co-authorship network, the APL value would be significantly smaller than the number of nodes in the network. A smaller value in a social network is believed to mean quicker spread of information and better communication between people. Our first important result showed that the APL of co-authorship networks is not directly correlated to the size of the network, instead it is affected by the following two factors:

- The coverage of the highest degree nodes.
- The size of the dataset.

The node removal experiments showed the different impacts on differently connected networks. In a network that the high degree nodes reach the most of other nodes in one step, such as a group-based network, the length of the APL almost tripled when the high degree nodes are removed. While in a network with less coverage by the high degree nodes, the APL is also extended significantly. In both cases, the high degree nodes play a key part in making the APL short.

For the APL affected by the size of dataset, our results showed that for a fixed pair of authors, their distance would be shorter in a larger dataset compared to a smaller dataset. This result appears contradictory to the fact that APL is not directly correlated with the size of network. But what this result actually means is the distance between a fixed pair may be shortened due to new nodes added to the network, but path length of the new pairs resulting from the additional nodes would offset the APL reduction.

The second important result of our investigation is the network metrics in this smaller-scale school-wide database are not significantly different from multi-institute domain-wide databases. The APL of the School of Electronic and Computer Science (ECS) community is 4.05, while the APL of the entire high energy physics community is 4.0. The collaborator number in our 1995-1999 dataset was calculated to be 2.68, slightly lower than the computer science 3.59. The metrics like collaborator number are invariant within the same dataset: it shows an entire network's collaboration size. But the APL does not have this invariance. Similar APLs represent the similar communication efficiency inside those particular networks. But the ratio between the number of papers and authors is very different. ECS researchers produced 18 papers per author while other databases only show below 2 papers per author in the same period. This indicates either the researchers in ECS are more productive than those in biomedical, physics or mathematics, or they are more scrupulous in their use of repositories. Large errors may have been introduced due to the difference between the methods used in counting the author numbers in this study and in other studies, and the incomplete data problem this study experienced.

Finally, the ECS's collaboration network follows the power law distribution. Most of the authors only have one or two collaborators, while most of the collaborations occur between only a few authors. This result is in line with Newman's work [17].

Social network analysis provides evidence for the connections between groups of individuals. It is these connections that provide channels for flow of information and the sharing of knowledge. In an environment such as a university, increasingly moving to interdisciplinary modes of research and funding, the existence of an efficient small-world network, covering its entire cohort of research-active individuals is crucial. This report shows that the ECS community is a small-world network that the knowledge-sharing is as effective as those communities formed by an entire discipline.

5.1 Critical Reflection

The ECS EPrints is not a perfect model of the entire literature that ECS authors have contributed to. Even with its imperfections, it shared common features with the data sources used by Newman.

In the path length analysis section, although using a school wide database has the advantage of knowing the researcher's group and role, it is limited in the sense that it cuts off those cross institutional links, resulting in an incomplete collaboration circle for the researchers. Many authors who do not have IDs in the database are from different institutions or different schools in Southampton.

In the node removal experiment, we used the degree threshold in selecting nodes to remove. As there were many nodes with the same degree, one iteration removed multiple nodes, resulting in wide gaps, as in figure 4.6 on page 26. A better solution would be to sort the nodes based on their degrees, and add one extra node according to the degree in each iteration.

Bibliography

[1] R. Albert, H. Jeong, and A.L. Barabasi. Diameter of the world wide web. *Nature*, 401 (6749):130–131, 1999.

[2] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614, 2002. ISSN 0378-4371. doi: DOI:10.1016/S0378-4371(02)00736-7. URL `http://www.sciencedirect.com/science/article/B6TVG-45S9HG2-1/2/dff30ba73ddd8820aca3e7f072aa7885`.

[3] Boutell.Com, 2009. URL `http://www.boutell.com/newfaq/misc/sizeofweb.html[Accessed14-9-2009]`.

[4] C. Chen. Trailblazing the literature of hypertext: author co-citation analysis (1989–1998). In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia: returning to our diverse roots: returning to our diverse roots*, pages 51–60. ACM New York, NY, USA, 1999.

[5] R.M. Fernandez, E.J. Castilla, and P. Moore. Social capital at work: Networks and employment at a phone center 1. *American Journal of Sociology*, 105(5):1288–1356, 2000.

[6] M. Granovetter. Economic action and social structure: the problem of embeddedness. *American journal of sociology*, 91(3):481, 1985.

[7] M.S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6): 1360, 1973.

[8] Web Science Research Initiative, 2009. URL `http://webscience.org/[Accessed23-9-2009]`.

[9] N. Jahn. The methodological status of co-authorship networks. 2008.

[10] P.D. Killworth and H.R. Bernard. The reversal small-world experiment. *Social networks*, 1(2):159–192, 1978.

[11] J.M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, 2000.

[12] Jon Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM. ISBN 1-58113-184-4. doi: http://doi.acm.org/10.1145/335305.335325.

[13] C. Korte and S. Milgram. Acquaintance networks between racial groups: Application of the small world method. *J. Personality and Social Psych*, 15:101, 1978.

[14] N. Lin and M. Dumin. Access to occupations through social ties. *Social Networks*, 8 (4):365–385, 1986.

[15] N. Lin, M. Walter, Ensel, and John C. Vaughn. Social resources and strength of ties: Structural factors in occupational status attainment. *American Sociological Review*, 46(4):393–405, 1981.

[16] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404, 2001.

[17] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):16131, 2001.

[18] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):16132, 2001.

[19] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, Jul 2001. doi: 10.1103/PhysRevE.64.025102.

[20] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[21] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(90001):5200–5205, 2004.

[22] M.E.J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *LECTURE NOTES IN PHYSICS-NEW YORK THEN BERLIN-*, 650: 337–370, 2004.

[23] JJ Potterat, RB Rothenberg, and SQ Muth. Network structural dynamics and infectious disease propagation. *International journal of STD & AIDS*, 10(3):182, 1999.

[24] E. Rosenthal. Social networks and team performance. *Team Performance Management*, 3(4):288, 1997.

[25] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[26] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, page 301, 1998.

[27] H.D. White and K.W. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998.

[28] Yang Yang, Au Yeung C, M Weal, and Davis H. The researcher social network: A social network based on metadata of scientific publications. 2009.

[29] N. Zekri and J.P. Clerc. Statistical and dynamical study of disease propagation in a small world network. *Physical Review E*, 64(5):56115, 2001.

Appendix A

Oversized Figures

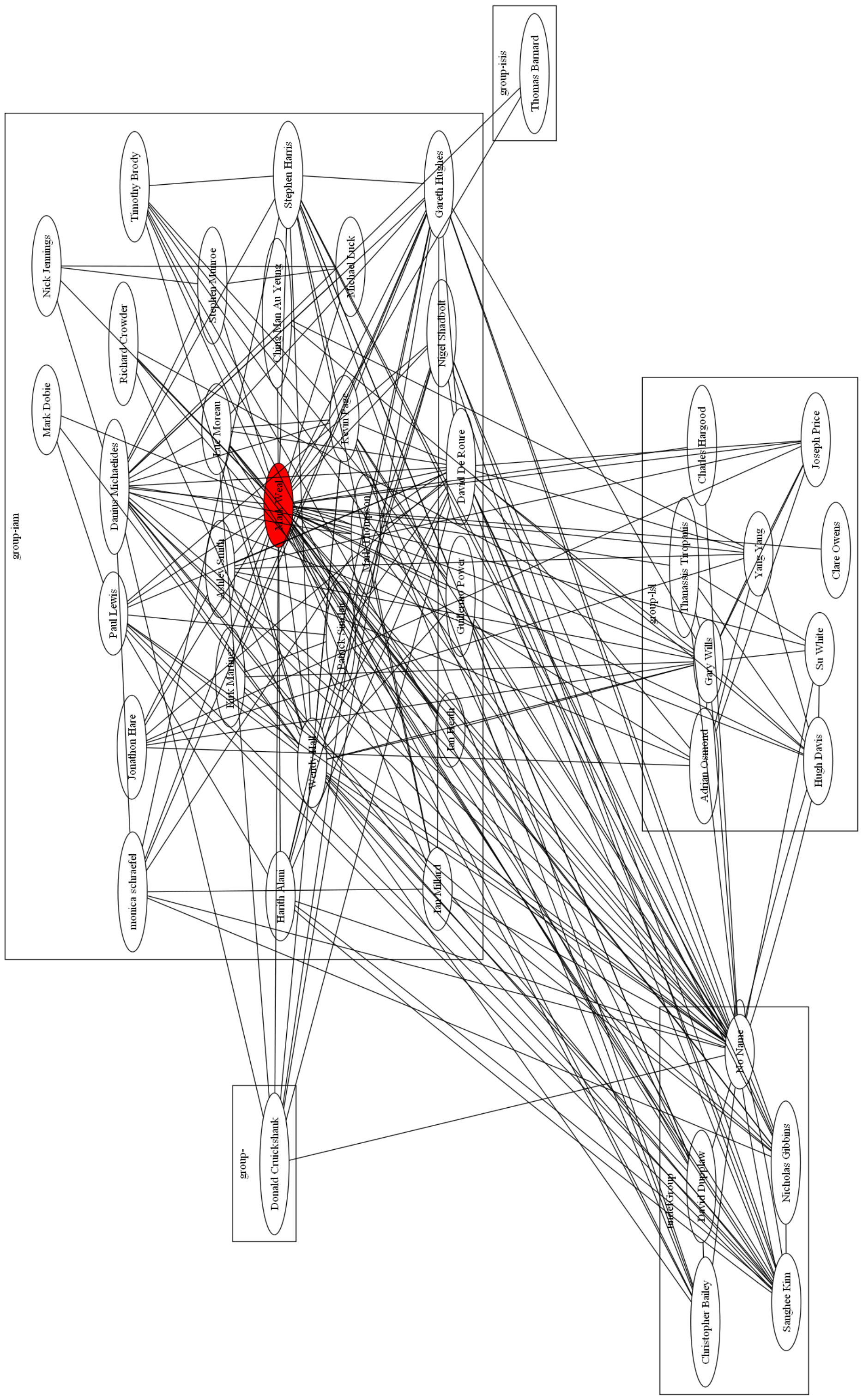
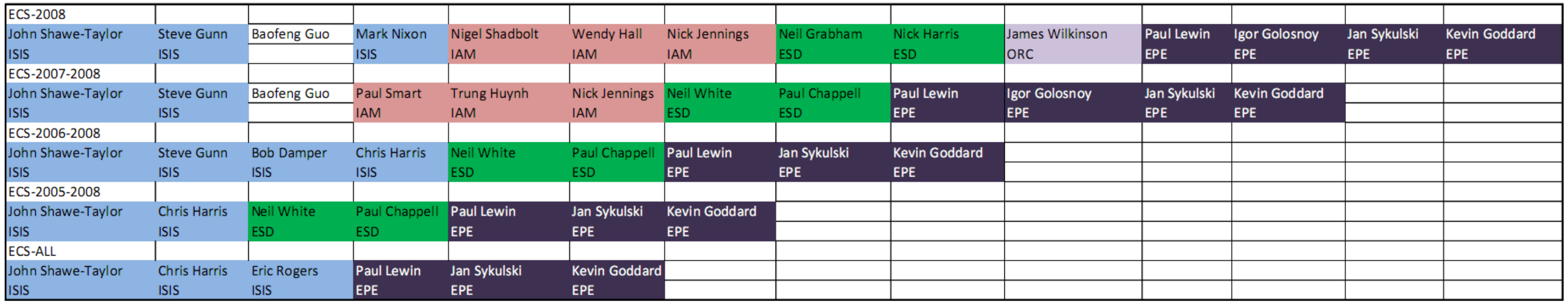


Figure A.1: Dr Mark Weal's collaboration network.



ECS-2008													
John Shawe-Taylor	Steve Gunn	Baofeng Guo	Mark Nixon	Nigel Shadbolt	Wendy Hall	Nick Jennings	Neil Grabham	Nick Harris	James Wilkinson	Paul Lewin	Igor Golosnoy	Jan Sykulski	Kevin Goddard
ISIS	ISIS		ISIS	IAM	IAM	IAM	ESD	ESD	ORC	EPE	EPE	EPE	EPE
ECS-2007-2008													
John Shawe-Taylor	Steve Gunn	Baofeng Guo	Paul Smart	Trung Huynh	Nick Jennings	Neil White	Paul Chappell	Paul Lewin	Igor Golosnoy	Jan Sykulski	Kevin Goddard		
ISIS	ISIS		IAM	IAM	IAM	ESD	ESD	EPE	EPE	EPE	EPE		
ECS-2006-2008													
John Shawe-Taylor	Steve Gunn	Bob Damper	Chris Harris	Neil White	Paul Chappell	Paul Lewin	Jan Sykulski	Kevin Goddard					
ISIS	ISIS	ISIS	ISIS	ESD	ESD	EPE	EPE	EPE					
ECS-2005-2008													
John Shawe-Taylor	Chris Harris	Neil White	Paul Chappell	Paul Lewin	Jan Sykulski	Kevin Goddard							
ISIS	ISIS	ESD	ESD	EPE	EPE	EPE							
ECS-ALL													
John Shawe-Taylor	Chris Harris	Eric Rogers	Paul Lewin	Jan Sykulski	Kevin Goddard								
ISIS	ISIS	ISIS	EPE	EPE	EPE								

Figure A.2: Path length evolution over time. A 13 step path has shortened to 5 steps when the dataset gets larger. The number of groups (colour represents group) involved in the path reduced from 5 to 2.

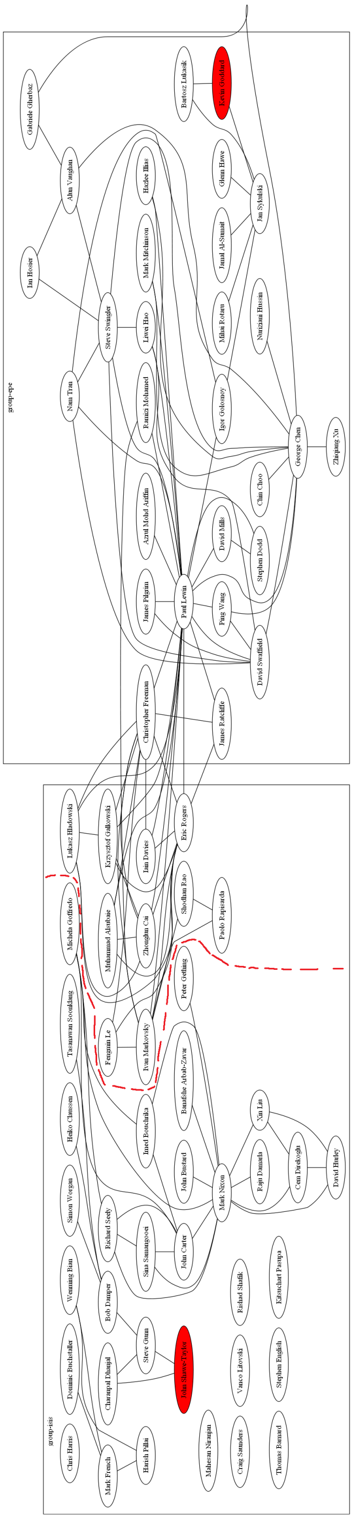


Figure A.3: The author collaboration between the ISIS and the EPE group in year 2008. Red nodes are the authors need to be connected. Red dotted line shows the split in the ISIS group

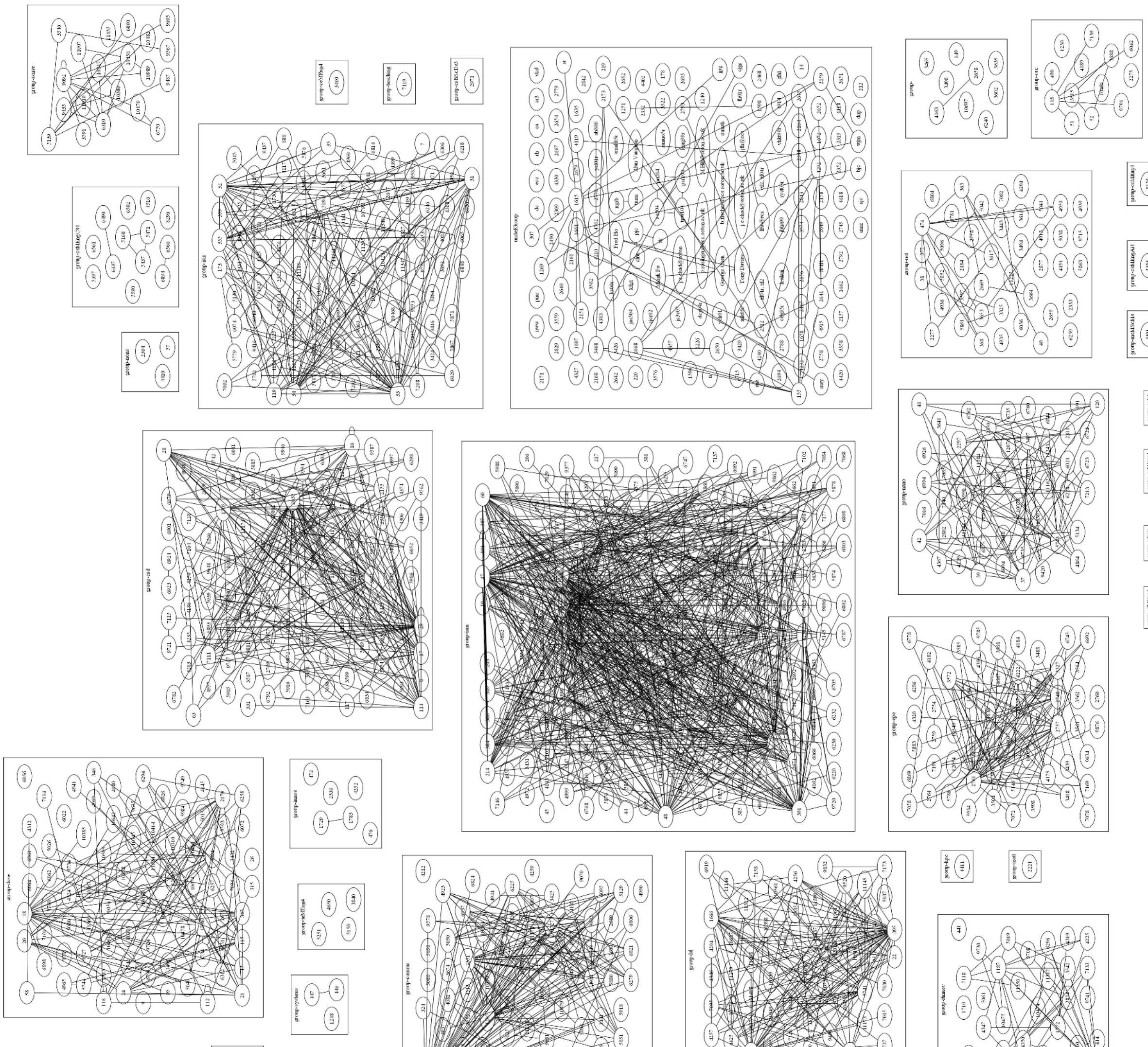


Figure A.4: Intra-group collaboration in ECS.

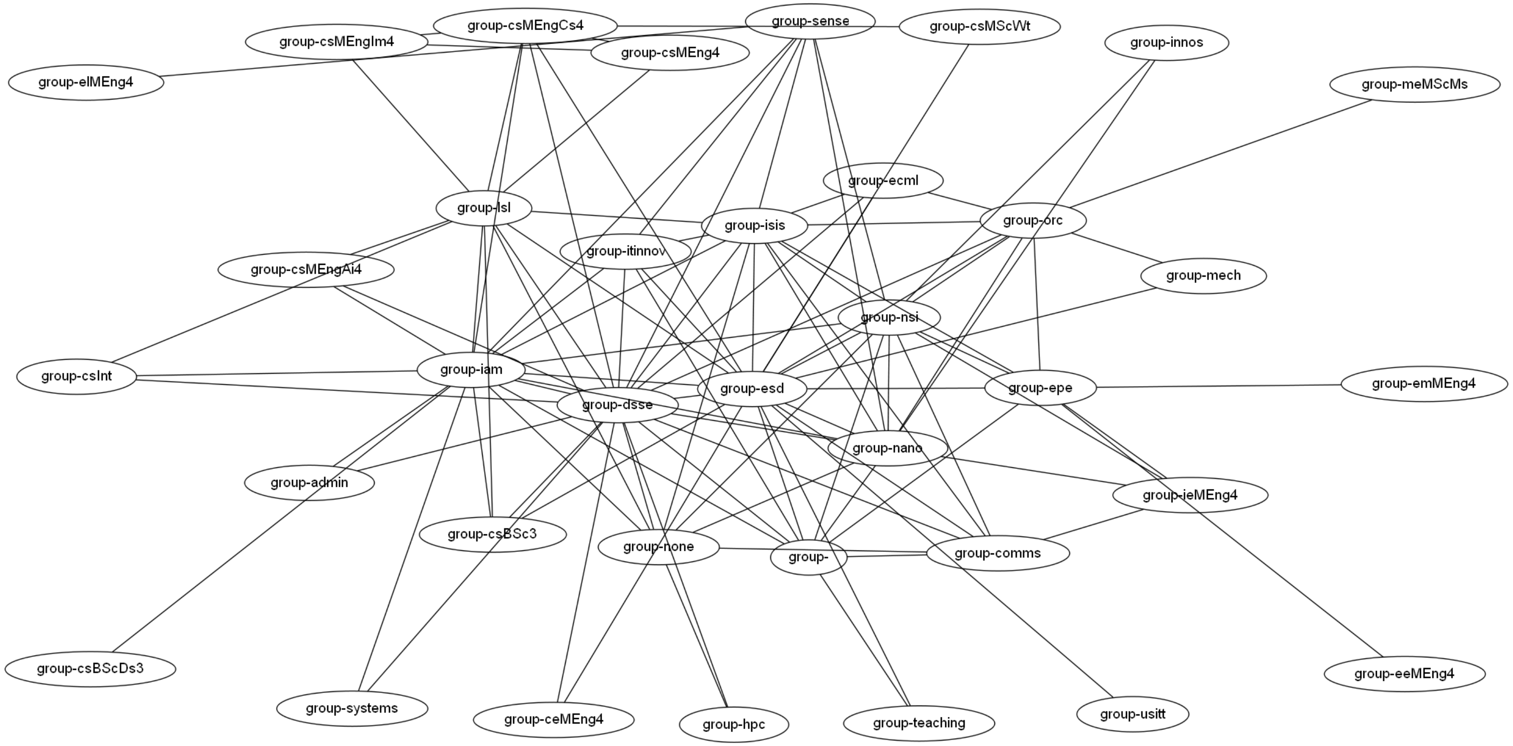


Figure A.5: The group level collaboration network in the ECS. The length of the edge represents the amount of author level collaborations between the groups

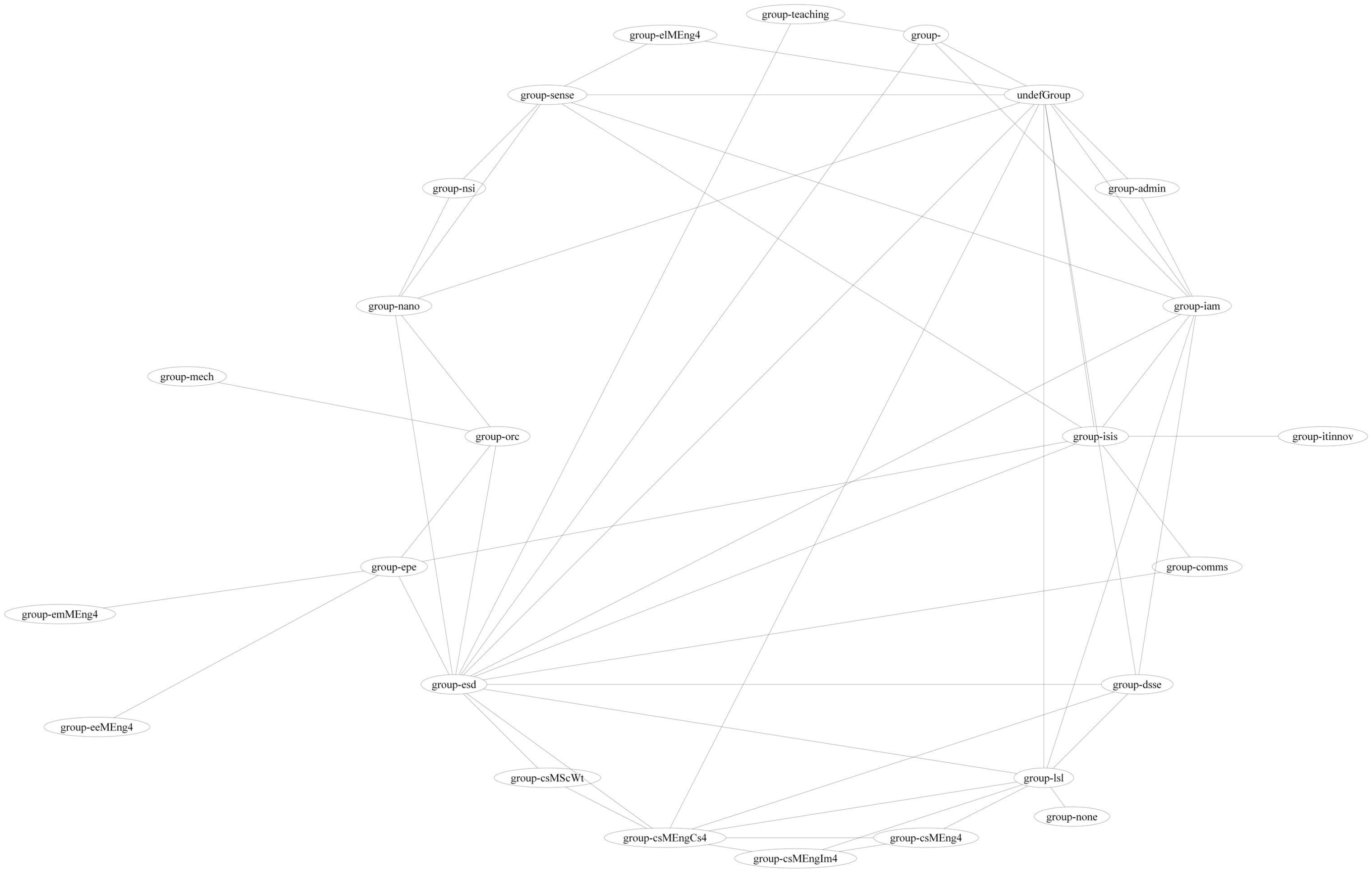


Figure A.6: The group level collaboration in 2008 in the ECS. This is a weakly connected graph considering each node represents dozens of nodes.

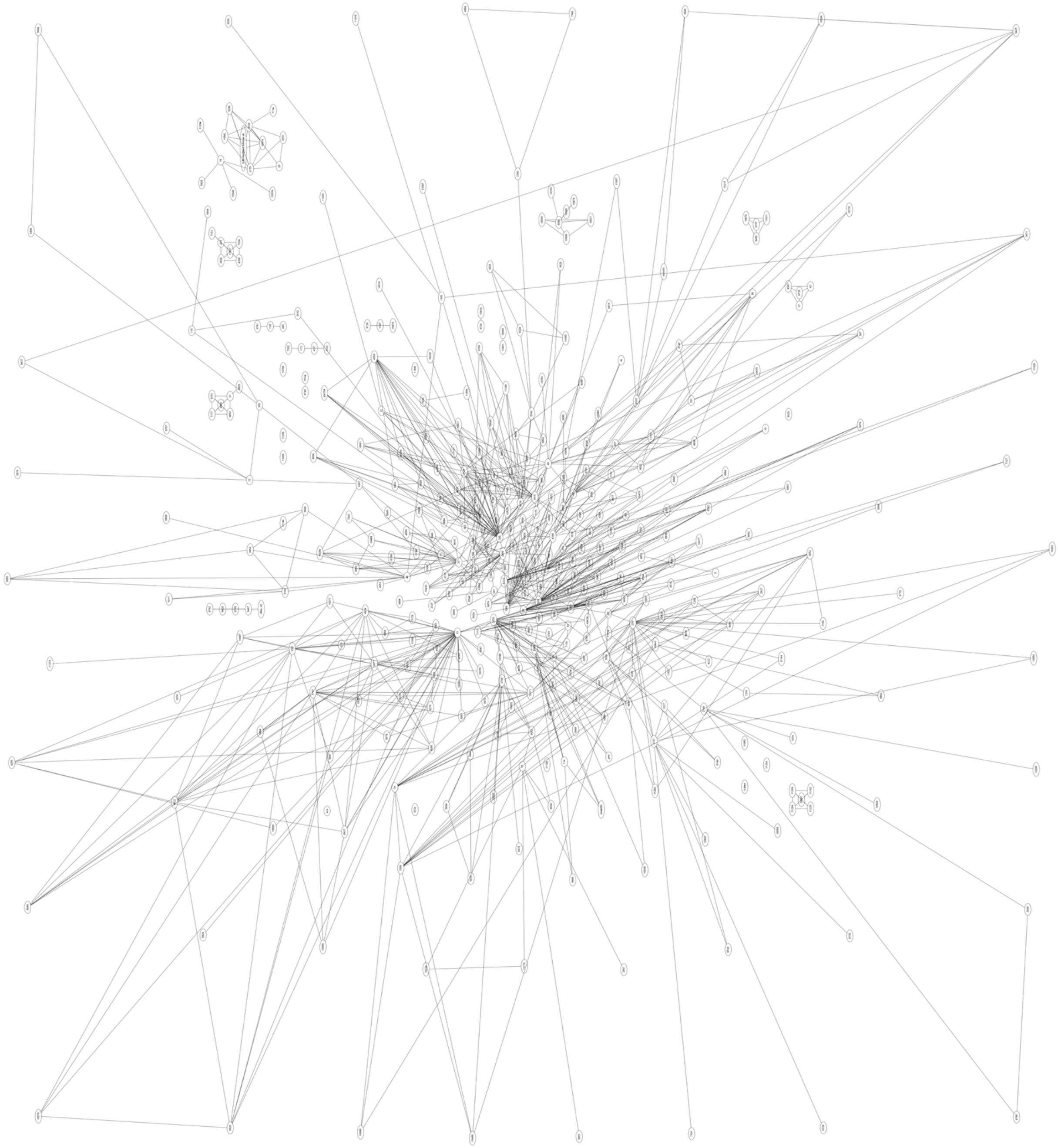


Figure A.7: Author collaborations in 2008

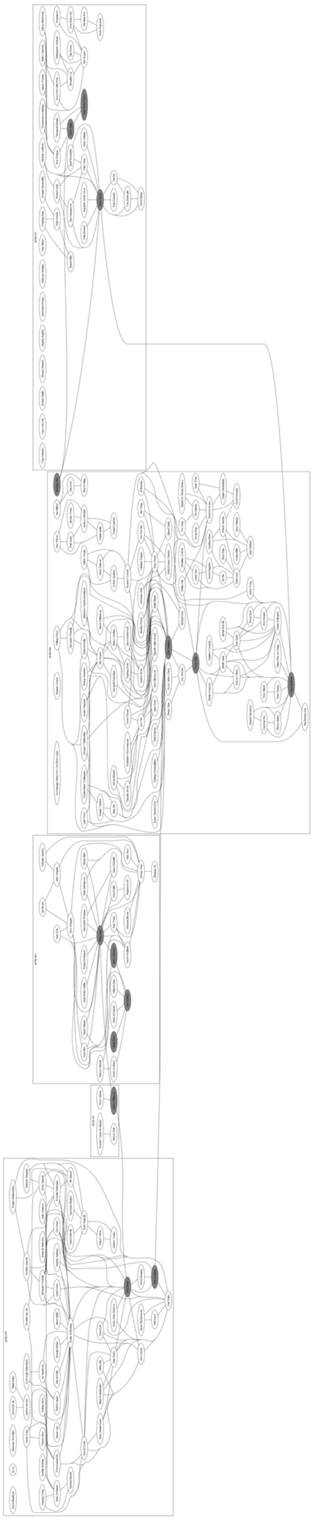


Figure A.8: Following a 13 step path in the 2008 dataset. Grey nodes mark out the path. Full sized figure is attached.

Appendix B

Project Plans

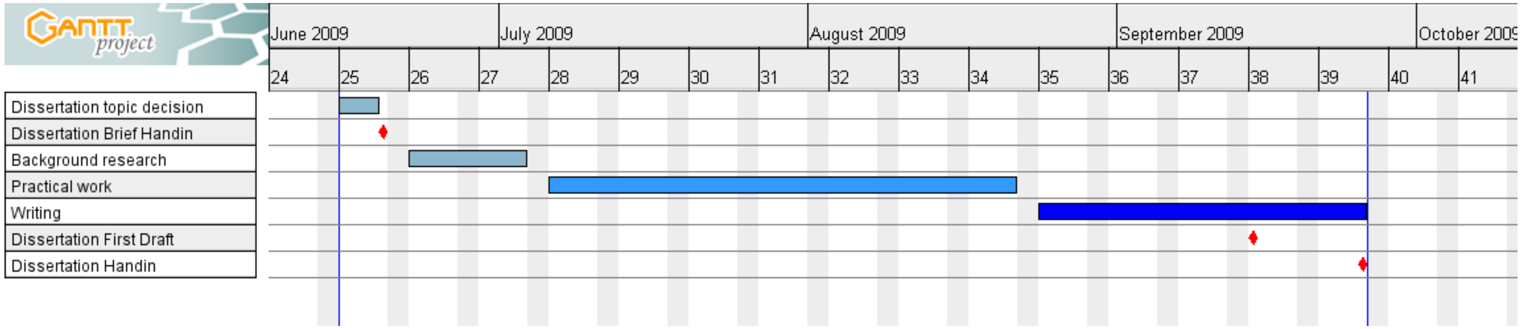


Figure B.1: Initial project plan made on the 12th of June. The official project start date is 15th of June, which is marked by left blue line. The project end date is 25th of September, marked by the right blue line. Red dots are various milestones and deadlines, different coloured bar represents different types of work. As the research topic is still not decided at this point, this is only a broad plan.

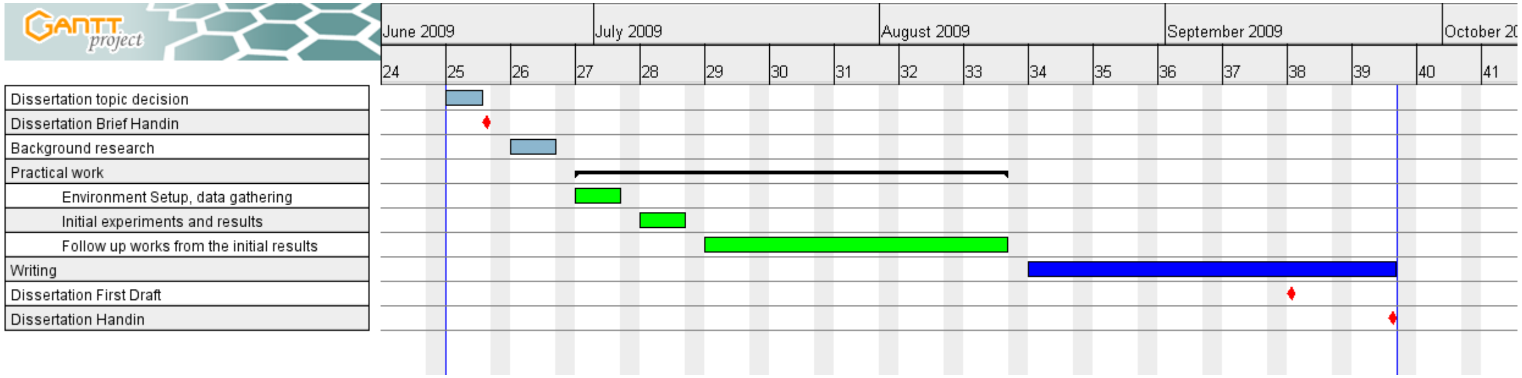


Figure B.2: The first review of the project plan made on 24th of June(half way into 26th week on the plan). After some background research, I moved the practical work plan forward a week. A more detailed practical work is also planned.

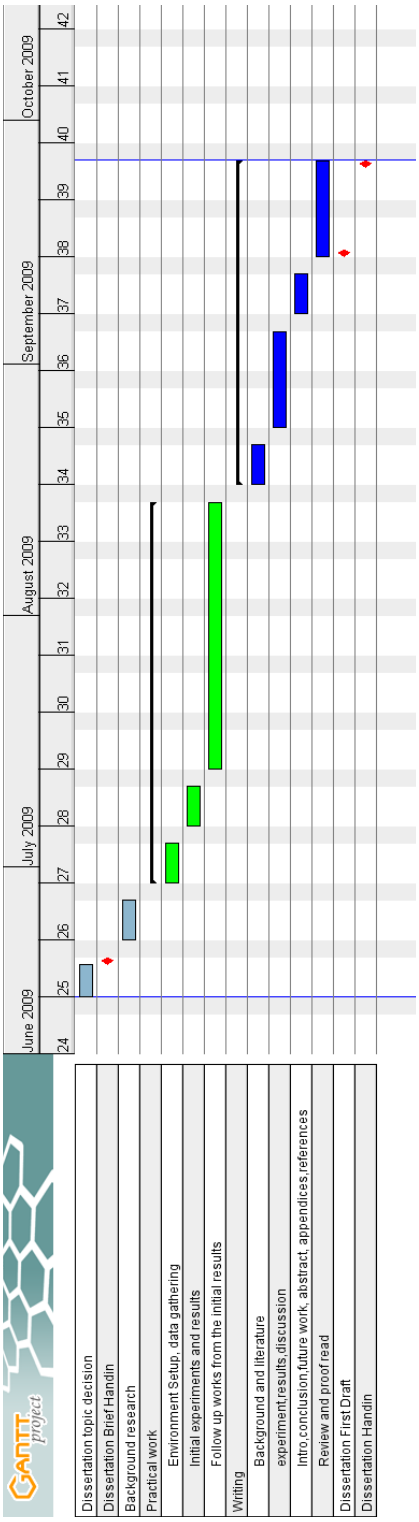


Figure B.3: The second review, made on the 5th of August. Added a more detailed writing plan

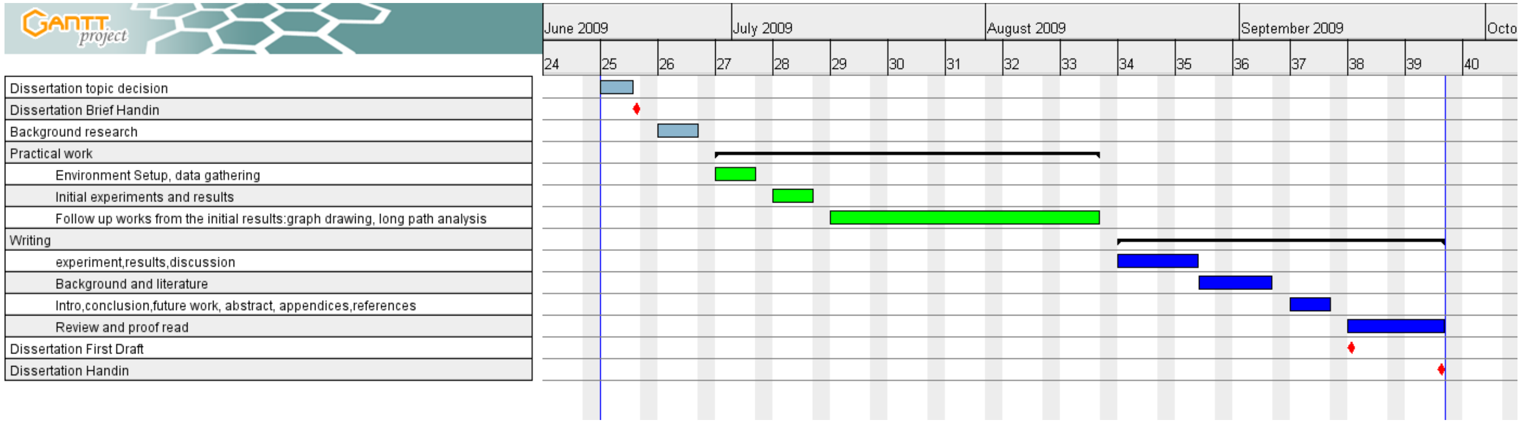


Figure B.4: The third review, made on the 20th of August. Some change made to the writing plan as starting to write the report.

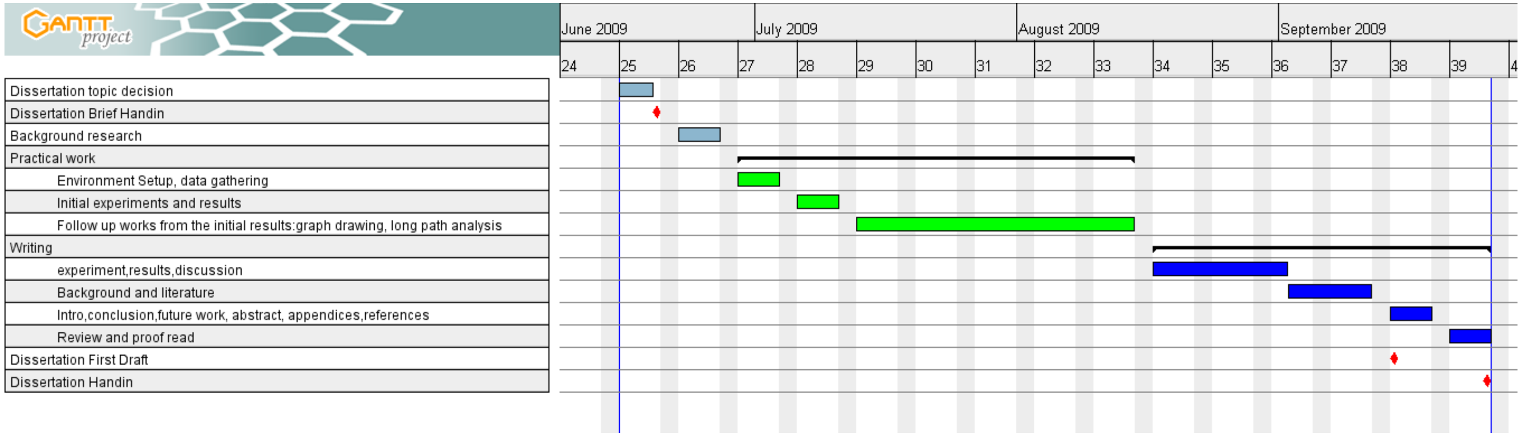


Figure B.5: The working progress of the project made on the 22nd of September. It shows the process of the project though this 3 month period, gives a good reference for planning a future project.