A Nine Month Progress Report on investigation of Social Network and Bibliometric Network

by

Jiadi Yao

University of Southampton
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science
September 2010

Contents

Chapter 1

Introduction

This is a nine month report detailing my initial research and experiments toward the understanding of the reciprocal relationship between social networks and bibliometric networks. This work is trying to answer some of the follow up questions in my previous MSc course project, which focused on analysing an institutional co-authorship network[67]. The research questions raised are split into two parts.

- To focus onto individuals: How does the co-authorship network reflect the actual working relations between the individuals? How does their job position co-relate to their network position?
- To scale up and include bigger entities: What does the universities' co-authorship network graph look like? Is there a split in co-authorship network between top universities and others? If so, what are the factors that split universities up?

This report attempts to rank the authors using the bibliometric network structure and analyses the university collaboration patterns.

1.1 Outline of the Report

Chapter 2 reviews previous works on social network and bibliometric network from various disciplines, giving an overview of the subject and providing a solid foundation which this work will be built on. It includes a review of existing software and tools for effectively conducting network analysis and graph visualisation for this work.

Chapter 3 describes one of the experiments that rank authors from ECS based on the co-authorship graph structure.

Chapter 4 describes the other experiment which analysed university level co-authorship collaboration using the ACM data.

Chapter 5 summarises the report and proposes future work.

Chapter 2

Literature Review

To fully understand networks attracted research from a wide range of fields including mathematics, physics, computer science and social science. Mathematicians consider networks as graphs. They use vertices and edges to describe a network, then by experimenting the way the vertices connected the by the edges, to understand the properties of networks; Physicists and Computer Scientists are interested in understanding how real networks are formed and evolve, for example computer networks, the transport network and the infectious disease transmission network. Social scientists are concerned with networks that involve people. They would like to know how people interact, how people obtain information and how people collaborate within their people-network. This class of networks that models the relationship between people are often called social networks.

In this chapter, we give an overview of research in various domains that concerning social networks and bibliometric networks.

Section 2.1 reviews network studies from the mathematical perspective, primarily the network models that are widely used in most network analyses.

Section 2.2 reviews works in the social science domain.

Section 2.3 reviews works in bibliometric analysis and how it relates to social networks.

Section 2.4 evaluates some common network analysis tools.

2.1 Network Models

The mathematical network model provides a solid foundation for network analysis. Any network can be modelled by a graph – nodes represent objects of interest and links represent relations between those objects. Modelling real networks using graphs recorded as early as 18th century. Leonardo Euler tackled the famous Königsberg's Seven Bridges problem – is it possible to traverse the seven bridges exactly once – by modelling the islands and bridge connection as a graph. He proved that the graph cannot be traversed only once.

A mathematical graph is defined as a pair of sets $G = \{V, E\}$, where V is set of vertices (or nodes) $v_1, v_2 ... v_n$ and E is set of edges (or links) that connect two vertices. The edges can also have values attached, so the graph becomes a valued graph. we discuss it in section 2.2.1.

In the following sections, three network models that were most widely studied are introduced.

2.1.1 Random Network Models

This is the class of network models that include the original Erdős and Rényi random graph model[20, 21] and the variations. The original model defines a very simple graph: a graph $\mathrm{G}_{n,p}$ is defined as n nodes then connects each pair of nodes with probability p. In fact, graph $\mathrm{G}_{n,p}$ is not a single graph, but the collection of graphs with n nodes and all the possible ways of connecting the nodes together with probability p.

There are two critical points that the Erdős and Rényi's graph model cannot model in social networks:

1. Degree distribution[1]. Due to the random nature of this model, the degree distribution follows the Poisson distribution. This is very different from many real networks, for example, social networks and citation networks follow the power-law degree distribution[44, 53].

2. Clustering. Social networks have high clustering[25], indicating a locally well connected structure. However, random network model can not produce this local structure due to its random nature.

There are variations of the Erdős and Rényi network model to address these problems. The configuration model[37] and Chung and Lu's model[11, 12] specifically targeted the degree distribution of random networks. Holland and Leinhardt [29], Strauss [22, 58] proposed models to address the clustering. But the common problem of these variations is that they become too complex to be useful in other studies.

So researchers start questioning themselves: are we starting from the correct foundation for modelling social networks?

2.1.2 The Scale-Free Network Model

The Erdős and Rényi random network model is one of the simplest yet most studied network models. However, as we have already discussed, it has major weaknesses in modelling social networks. Barabási and Albert [3, 4] presented a new way of modelling networks. They emphasized on the growth of networks found in real life. The social network, the citation

[1] Degree of a node is the number of edges connected to that node.

network and the World Wide Web are evolving networks. They all started with few nodes, then new nodes created and attached to existing nodes in the graph, finally resulted the current network. Barabási and Albert showed that in order to produce a real network's degree distribution, whenever a new node is added to the graph, the node must have a higher chance to connect to nodes that already have many connections. For instance in citation network growth, a new publication has higher chance to cite one that thousands of other publications cite, than one with only a few dozen publications cite.[2] They call this the 'preferential attachment'. The resulting topology is that their degree distribution follows a power law. This means that most nodes have very few links, but the remaining few nodes have all the rest.

The importance of their contribution is not only on a new network model, but also a whole new way of viewing a network – it is a dynamic, evolving structure. Some of the network features are rooted in evolution of the networks rather than the network topological characteristics.

2.1.3 Small-World Phenomenon

Before we talk about this phenomenon, we need to introduce a metric that measures the connectedness of a network – the Average Path Length (APL). The APL in a network is the average of the shortest path between all pairs of nodes. A network with APL of 3 tells us that on average, the path length between *any* pair of nodes is 3. The Small-World Phenomenon that often found in many real networks is the observation that a large network – with millions or billions of nodes – has only a small APL. The human acquaintanceship network [60] consists of billions of nodes and its APL is only 6; the mathematics co-authorship network with 250 thousand researchers[45] has an APL of only 7. One of the properties of this class of networks is that information transmitted on them is much faster than, for example, a network that has APL in the order of thousands or millions. Therefore, the small-world effect is desirable in networks like the scientific collaboration network and the WWW, where knowledge can channel through quickly, but not so desirable in situations like disease transmission or the spread of rumours in social networks. More recently, the small-world phenomenon has taken a precise meaning: networks are said to show small-world phenomenon if the APL of the network scales logarithmically or slower with the network size for fixed average degree.

Another property of real-life small-world networks is that nodes are locally clustered. This means that if node A is connected to node B and C, then B and C are very likely connected too. One demonstration in social networks is two close friends of someone are very likely friends themselves too.

The Erdős and Rényi random network model and its variations reproduces the small-world effect well[5, 12, 18, 23]. But as we have already discussed in section 2.1.1, the random

[2] This can be simply explained by probability: If thousands of publications cite a paper, then this paper is much easier to be found than one that only a few paper cites.

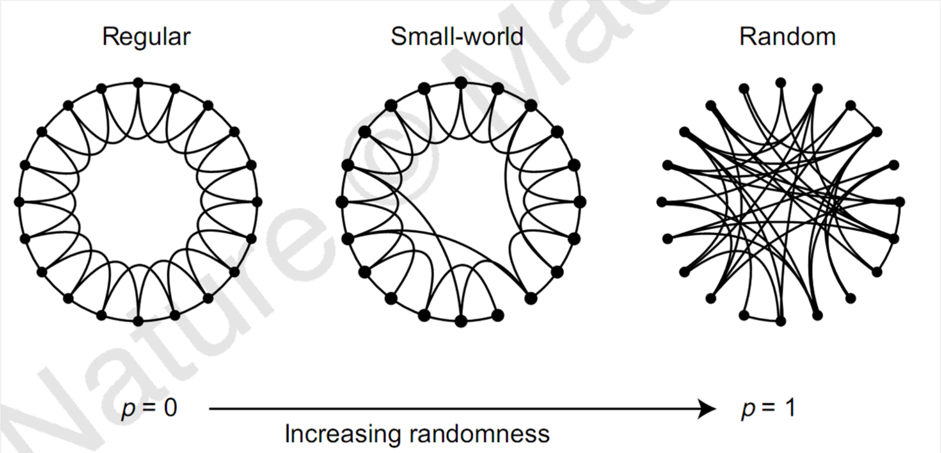


Figure 2.1: Watts and Strogatz Model. Left: the regular ring lattice with no randomness; middle, some randomness introduced when connecting neighbours, the network became small-world; right, a complete random graph.

network model cannot easily produce high clustering.

Watts and Strogatz[62] proposed a simple model that has both small-world effect and high clustering (Figure 2.1). The model starts with a ring of nodes, then each node is connected to one nearest neighbour from both sides. A randomness parameter p is introduced, such that the amount of edges in the model is randomly rewired according to p. When $p = 0$, no edge is rewired and the graph remains a lattice; while when $p = 1$ the graph is completely rewired and becomes a random graph. By varying the randomness p, there is a sizeable region, as shown by Watts and Strogatz using numerical simulation, where the model has small-world phenomenon and is highly clustered.

This model demonstrated observations in many social systems, where most people are friends with people they are geographically close to – colleagues, house neighbours, classmates – and the lattice represents these connections. Many people also have a few friends that live a long way away – friends live in other cities or other countries – the randomness adds long distance connections to the network.

Analysis of this model shows surprising result: in order to convert a lattice network into a small-world network, only a tiny fraction of rewiring is required. What this means is that the small-world phenomenon found in social networks is stable and is not on the edge of collapse.

There are many variations of the Watts and Strogatz model. A much studied variant was proposed by Newman and Watt [46], which randomly adds edges to the graph but does not remove edges from the regular lattice. This prevents isolated clusters forming so that it is easier to analyse. Models with higher dimensions have also been proposed and studied[16, 39, 47, 50], and the results are qualitatively similar to the one-dimensional case.

2.2 Social Networks From A Social Science Perspective

2.2.1 Tie Strength

There is a thread of study in social science that considers the tie strength between people. These studies demonstrated that the strength of the relationship plays vital roles in shaping the social structure. Granovetter[26] conducted an experiment on how people acquire information about job opportunities in a small town. He found that their social contacts are the primary channel. But what is also interesting is that most of these social contacts were not even their friends, they were just acquaintances. So he suggested that 'weak ties' between people are where they get valuable information form. His explanation was that even though family members and friends would like to help when they are between jobs, because they and their friends are a tightly connected network they don't have much new information to offer. However acquaintances would have access to information from their side of the circle of friends, providing valuable opportunities.

There are many follow-up studies: Onnela et al.[49] studied the telephone communication network. They used the length of the telephone call as a measure of the strength of the tie. Studies on Facebook and Twitter using message frequency for tie strength are also conducted [30, 35]. Ahuja et al.[1] also proposed methods to use tie strength to partition networks.

2.2.2 Social Data Collection

When conducting social network analysis, it is important to understand what kinds of data that need to be collected for an effective analysis. Scott[8] produced a social network analysis handbook detailing common methodologies. He introduced two principal types of data one needs to collect to conduct an analysis: 'attribute data' and 'relational data'. Attribute data is regarded as opinions, attributes and behaviours of individuals. These are properties, qualities or characteristics that *belong to* the people in study. Relational data are the contacts, ties and connections that *relate* one individual to another. Relational data cannot be further reduced to properties of the individuals themselves. This formalisation provides a good guideline in the data collection stage of any network analysis.

In this study, the social data is encapsulated in the research paper's metadata. The author's name and affiliation are the attribute data that used to describe the author; whenever two authors appear on a same paper, it implies a relation between the two authors.

Computational Approaches The data collection methods used in many social science studies were mostly questionnaires. The downside is that the data is subjective to the person filling out the questionnaire. For instance, different people may give different definitions to friendship. In addition, the amount of data one could collect by handout questionnaires are rather limited.

In recent years, as more and more social interactions are moving to the web, social scientists are starting to obtain data from online resources. King et al. [32] grouped on-line social data sources into the following categories:

- Social Networks. eg. Facebook, where people are connected in this virtual society.
- Social Media. eg. Youtube, where people are connected because they have viewed, or commented on the same media.
- Social Game/Human Computation. People are connected because they interacted with each other in a game.
- Social Bookmarking/Tagging. eg. Del.icio.us, where people are connected because they have bookmarked the same resource, or used the same word to tag resources.
- Social News and Social Knowledge Sharing. eg. Wikipedia, where people get connected because they edited the same article together.

2.3 Social Network Meets Bibliometric Network

2.3.1 The Citation Network

Citation networks are classic knowledge networks. Papers represent nodes and citing papers represent directed links between papers. Because only the later papers can cite the previous published papers, this is a non-cyclic network and arrows on the links point back in time. Researchers started studying this network in the 1950s. Price[52] investigated the patterns of the citation network. He found that the small amount of papers are cited more frequently than average, while the majority of papers are cited less frequent than average. This work was one of the first empirical results for scale-free networks introduced in section 2.1.2.

Citation data is often used as an indicator to measure scientific performance. Redner[54] studied the citation count and scientific impact; Cronin[14] investigated the h-index and ranking of authors; Cole[13] and van Raan [61] evaluated the influence by awards, honours, and Nobel laureateships.

However, other studies suggested that citations are not a suitable measure for scientific activities [24, 66]. They claim that the citation depends on many factors besides scientific impact. These include, for example,

- Time-dependent factor: the more frequently a publication was cited, the more frequently it will be cited in the future [9, 52];
- Availability of the publication: physical accessibility [57], open access of publications[7, 27] and the publishing media influence the probability of citations[56];

- Author-reader dependent factors: results from Mählck and Persson [34] and White[63] showed that citations are affected by social networks, authors cite personally acquainted author more often.

Bornmann and Daniel[6] reviewed the citing behaviour of scientists. They claim that at the micro-level, citing is a social and psychological process that is mixed with personal bias and social pressures; but at the macro-level, scientists give credits to colleagues by citing their work. Thus, citations represent intellectual or cognitive influence on scientific work.

Studies of the citation network enabled us to understand the structure of knowledge and anticipate developments in various domains. The network constructed using citation relations between published papers are generally treated as a knowledge network. Since papers are written by researchers, a network of researchers can also be constructed. In coming sections, we show how data gathered from publications can construct a network that represents meaningful social relations.

2.3.2 The Co-citation Network

Two types of co-citation networks are discussed commonly in the literature: the paper co-citation network and the author co-citation network. These are two different networks - one connects papers together and one connects people. We discuss the author co-citation network here.

The author co-citation network is generally defined as follows: a pair of authors is said to be co-cited if they are cited together in a later paper, regardless of which of their work is cited. In a broad sense, the researchers cite two papers together when the content of those papers are somewhat relevant. So previous authors who wrote those co-cited papers can have similarities in their research. The more the authors are co-cited, the stronger the similarity in their research work. We can consider citing a pair of paper together as one vote to 'these papers are similar', so the co-citation network represents a collective decision of later authors in grouping previous papers.

The idea of co-citation is an attempt to push the knowledge network onto the people who produced the knowledge.

The original methodology by White [64] only considers the first author of any given paper and disregards the contributions of other co-authors. Follow up studies by Persson[51] and Zhao[68] consider all authors listed on a paper, helping to identify the domains of authors who are seldom listed as first authors. Beyond that, Su[59] proposed an algorithm to discover authors with multiple expertise in a co-citation network.

A few domain co-citation analyses have also been performed. White et al. [65] analysed the information science field from 1972 to 1995 using the author co-citation. They generated maps of the top 100 authors in the field and used factor analysis to identify major specialities. They found that information science consists of two major specialities with little overlap.

Chen and Carr [10] used ACM publication data to study the structure of the hypertext literature. Authors cited less than 5 times during the period 1989-1998 were filtered, resulting in 367 authors. An author co-citation matrix was constructed and Principal Component Analysis (PCA) was applied. The temporal information of the papers was included in the visualization methods, allowing them to identify emerging research directions in the field.

Co-citation networks start to bring people to the network, and provide a unique author-knowledge network that allows us to understand the focus of papers, and hence individual author's research interests. In the next section, we look at more direct people-to-people relations – the co-authorship network.

2.3.3 The Co-authorship Network

In these networks two researchers form co-authorship links if their names both appear on the same paper, so the nodes are authors and links represent co-authorship. In research activities, co-authors generally know each other and many of them collaborate with each other. As a result, the co-authorship network is, to certain degree, a researchers' collaboration circle. While the networks discussed previously were not collaboration networks between people, the co-authorship network is truly a proxy to the social collaboration network of the researchers. It has attracted much research attention in recent years.

The Erdős Number Calculating the Erdős Number is one of the earliest co-authorship activities. The Erdős Number is a measurement of the number of collaboration hops a researcher co-authored with the famous Mathematician - Paul Erdős. Researchers who co-authored a paper with Paul Erdős have Erdős Number 1; researchers who co-authored a paper with a co-author of Paul Erdős have Erdős Number 2 and so on. Those authors who never co-authored a paper with Paul Erdős don't have an Erdős Number or are said to have an infinite Erdős Number. De Castro and Grossman[15] found that many famous researchers, whatever their research areas, have a finite Erdős Number. Because the famous researchers also are tightly connected with their own research domain, so it leads to the fact that the entire research community is connected through co-authorship. The implication is that the scientific research is a collaborative work rather than individuals making their own discoveries. Another finding by them, probably the obvious one, is that in order to have a smaller Erdős Number, whom one collaborated with is more important than the number of collaborators – quality is more significant than quantity.

Domain analyses Co-authorship analysis is widely used to understand publication and collaboration patterns among researchers in a specific domain.

Newman [40–42, 45] carried out a series of co-authorship analyses in 2001. He answered a broad variety of questions about collaboration patterns by analysing co-authorship networks, such as the number of papers authors write, how many people they write them with, the typical distance between researchers through the network. He compared these attributes across several domains – Biology, Physics, Computer Science and Maths. Here are some of his findings:

- The number of papers written per author is similar across the domains in the study;
- The number of authors per paper and the average number of collaborators vary substantially across domains;
- All of the subject domains have a largest component connecting at least 80% of the researchers;
- The average collaboration distance is small, typically 4 to 6 steps for a network containing millions of nodes;
- The clustering coefficient is much smaller than a random network expected value.

After Newman's work, Moody[38] investigated social science collaboration networks; Liu[33], Sharma[55] studied digital library community, and Elmacioglu[19] analysed database community using co-authorship network.

All of these works came to one similar conclusion – researchers are mostly connected, the distance between them is short and the network is highly clustered. The co-authorship networks, therefore, are small-world networks.

Network Dynamics The co-authorship relation is also used in many network evolution studies. Each publication has a published date, which is used as the time variable for a dynamic network.

Barabási[3] studied the co-authorship network and proposed a simple model that captured the network's time evolution. By studying the model, they discovered that the measurements on incomplete databases could offer opposite trends. For example, the node separation exhibits a decreasing tendency on datasets that only cover certain periods, while their numerical simulation uncovers this inconsistency. They also found that the average degree, the diameter and the clustering coefficient, which are commonly used to characterise a network, are in fact time dependent, therefore can not be used to characterise an evolving network. On the other hand, they claim that degree distribution is a stationary measurement for a dynamic network.

Newman[43] analysed what affects researchers' choice over who to collaborate with. He found that the probability a pair of researchers would collaborate with each other increases with the number of other collaborators they have in common; and the probability of a particular researcher acquiring new collaborators increases with the number of his or her past collaborators. This result demonstrated *preferential attachment* in co-authorship networks when one is deciding who to collaborate with.

Author Name Disambiguation There is a common problem in publication co-authorship: the names printed on the paper do not globally identify an author. Multiple authors may be mixed into one when their names are not spelt out in full or one author is recognised as several because different name spellings used on publications. The name ambiguity problem affects the accuracy of co-authorship network analysis and many applications that rely on unambiguous author names. As shown by Kang et al.[31] and On[48], the co-authorship network itself could help identify authors. The assumption is that an author would work closely with the same group of researchers. By analysing the frequent co-authors of a given name string, it is possible to recognise whether this particular name presents many authors or just one.

As we have shown here, co-authorship is a fruitful data source for social network analysis. It revealed authors's collaboration trends both within domains and across domains[38, 40–42, 45, 55]; it identified top authors and their affiliations as well as the pattern of co-authorship among them[33, 41]; it also provided data for studying the network dynamics in social networks[3, 43]. In the remaining of the PhD program, co-authorship data is used as the building blocks for researcher collaborations.

2.4 Evaluation of Network Analysis Tools

A powerful network-analysis tool would enable one to conduct network analysis effectively. This section reviews some common network analysis software. We use a network consisting of 1.6 million edges as the test network to assess whether the tool is able to handle large networks.

Network Workbench Tool

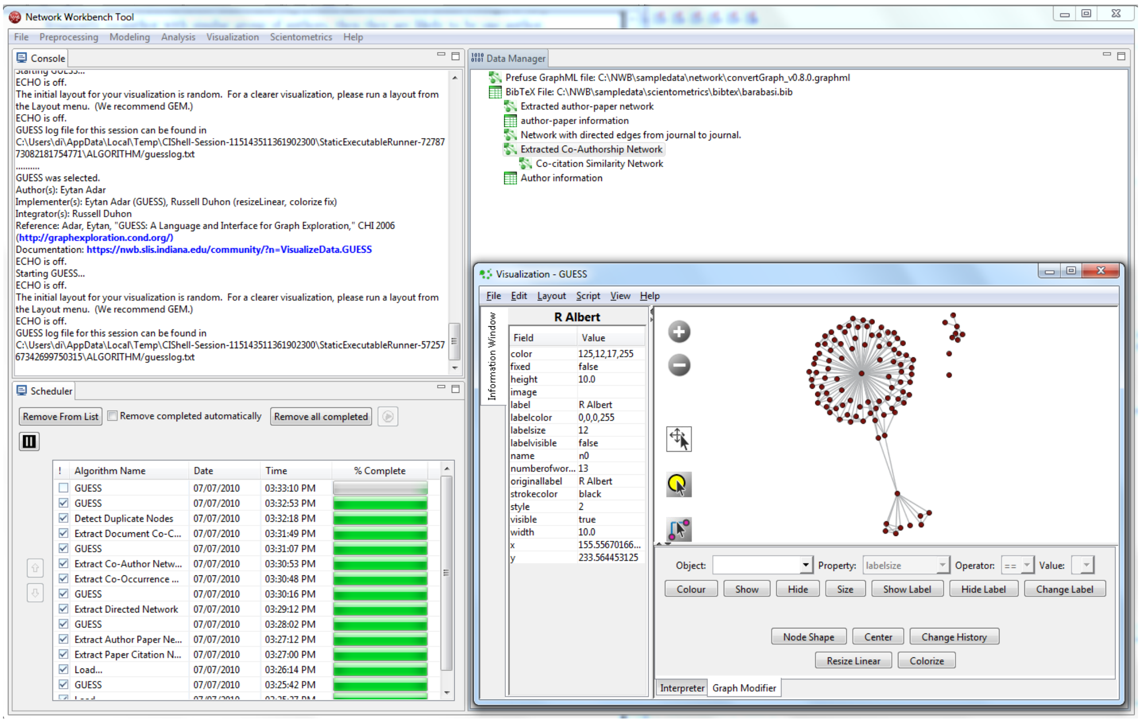
http://nwb.slis.indiana.edu/

The Network Workbench Tool is an open source project led by researchers from Indiana University and Northeastern University funded by cyberinfrastructure for network science center. It is network analysis software with graphical user interface(GUI) and it runs on most common desktop operating systems and supports a variety of network file formats.

This software provides a range of graph pre-processors, for example, extract nodes above or below certain criteria; remove self loops; delete high degree nodes – allowing one to trim the network before applying any algorithm. It computes most network properties by a single click and there are many pre-programmed algorithms ready to be applied to networks including PageRank, Hits, clustering coefficient, degree distribution and shortest path distribution. It also provides a GUESS module for graph visualisation.

It successfully loaded the test network and performed the analysis, but it was unstable when processing large networks.

This is the main tool that we used to analyse our networks.



```
Starting GUESS...
ECHO is off.
The initial layout for your visualization is random.  For a clearer visualization, please run a layout from
the Layout menu.  (We recommend GEM.)
ECHO is off.
GUESS log file for this session can be found in
C:\Users\di\AppData\Local\Temp\CIShell-Session-115143511361902300\StaticExecutableRunner-72787
73082181754771\ALGORITHM/guesslog.txt
..........
GUESS was selected.
Author(s): Eytan Adar
Implementer(s): Eytan Adar (GUESS), Russell Duhon (resizeLinear, colorize fix)
Integrator(s): Russell Duhon
Reference: Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006
(http://graphexploration.cond.org/)
Documentation: https://nwb.slis.indiana.edu/community/?n=VisualizeData.GUESS
ECHO is off.
Starting GUESS...
ECHO is off.
The initial layout for your visualization is random.  For a clearer visualization, please run a layout from
the Layout menu.  (We recommend GEM.)
ECHO is off.
GUESS log file for this session can be found in
C:\Users\di\AppData\Local\Temp\CIShell-Session-115143511361902300\StaticExecutableRunner-57257
67342699750315\ALGORITHM/guesslog.txt
```

Remove From List ☐ Remove completed automatically Remove all completed

!	Algorithm Name	Date	Time	% Complete
☐	GUESS	07/07/2010	03:33:10 PM	
☑	GUESS	07/07/2010	03:32:53 PM	🟩
☑	Detect Duplicate Nodes	07/07/2010	03:32:18 PM	🟩
☑	Extract Document Co-C...	07/07/2010	03:31:49 PM	🟩
☑	GUESS	07/07/2010	03:31:07 PM	🟩
☑	Extract Co-Author Netw...	07/07/2010	03:30:53 PM	🟩
☑	Extract Co-Occurrence ...	07/07/2010	03:30:48 PM	🟩
☑	GUESS	07/07/2010	03:30:16 PM	🟩
☑	Extract Directed Network	07/07/2010	03:29:12 PM	🟩
☑	GUESS	07/07/2010	03:28:02 PM	🟩
☑	Extract Author Paper Ne...	07/07/2010	03:27:12 PM	🟩
☑	Extract Paper Citation N...	07/07/2010	03:27:00 PM	🟩
☑	Load...	07/07/2010	03:26:14 PM	🟩
☑	GUESS	07/07/2010	03:25:42 PM	🟩

Prefuse GraphML file: C:\NWB\sampledata\network\convertGraph_v0.8.0.graphml
BibTeX File: C:\NWB\sampledata\scientometrics\bibtex\barabasi.bib
Extracted author-paper network
author-paper information
Network with directed edges from journal to journal.
Extracted Co-Authorship Network
Co-citation Similarity Network
Author information

R Albert

Field	Value
color	125,12,17,255
fixed	false
height	10.0
image	
label	R Albert
labelcolor	0,0,0,255
labelsize	12
labelvisible	false
name	n0
numberofwor...	13
originallabel	R Albert
strokecolor	black
style	2
visible	true
width	10.0
x	155.55670166...
y	233.564453125

Figure 2.2: GUI Interface of the tool with network visualisation

NodeXL

http://nodexl.codeplex.com/

NodeXL is a Microsoft Excel 2007 plug-in that enables the Excel to manipulate and analyse networks.

By leveraging the existing Excel interface and spreadsheet functions it means that it is easy and quick to start working. It does not load any established network file format, instead it recognises the vertex pair as an edge in the network. A few network pre-processing algorithms are available, for example, to convert duplicated edges into edge weight; extract vertices from edge list. Many other pre-processing operations can also be achieved using spreadsheet formulas. The calculation of basic network properties, the graph visualisation and the chart generation are also supported.

However, the network size this plug-in can handle is limited. Excel permits a maximum 1,048,576 rows in a workbook, hence any graph can have a maximum of that number of edges, which is below our test network size. Availability of this package is also limited as this plug-in only works for Excel 2007.

	GUI	OS	Loads Large Network	Visualisation	Extensible
NWB	✓	All	✓	✓	✓
NodeXL	✓	Windows	✗	✓	✗
SN Visualiser	✓	All	✗	✓	✗
SONIVIS	✓	Windows Linux	✗	✓	✗
Pajek	✓	Windows	✗	✓	✗
UCINET	✓	Windows	✗	✓	✗
Gephi	✓	All	✗	✓	✓
prefuse	✗	All	N/A	✓	✓
JUNG	✗	All	N/A	✓	✓
NetworkX	✗	All	N/A	✓	✓
SNAP	✗	All	N/A	✓	✓
igraph	✗	All	N/A	✓	✓

Table 2.1: Network Tools Comparison

Social Networks Visualizer

http://socnetv.sourceforge.net/

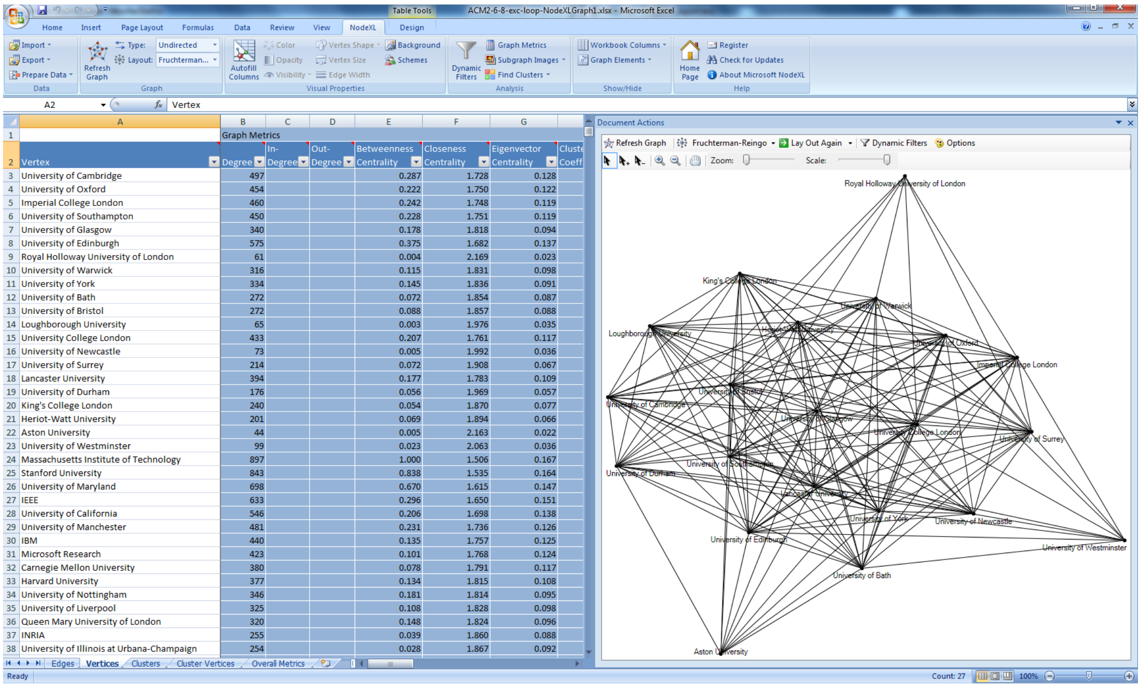
Social Networks Visualizer is an analysis and visualisation tool for social networks. It supports many graph file formats including Graphml, which we use heavily. One feature of this software is that it has an integrated web crawler to enable quick web graph analysis.

But Social Networks Visualizer doesn't provide a way to manipulate graphs and it treats any graph as directed, which limits its usage. It does not have a flexible graph visualiser and it failed our large network test.

SONIVIS

http://www.sonivis.org/

This is an open source project which aims to create network analysis and network mining software. Its graphical user interface is based on Eclipse. Like many other tools, it is able to calculate many graph attributes and supports graph visualisation. Unfortunately, it is still in development (as of July 2010), and it currently only supports direct database connection to import network data. Therefore we have no way to import our test data.



	A	B	C	D	E	F	G
1		Graph Metrics					
2	Vertex	Degree	In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
3	University of Cambridge	497			0.287	1.728	0.128
4	University of Oxford	454			0.222	1.750	0.122
5	Imperial College London	460			0.242	1.748	0.119
6	University of Southampton	450			0.228	1.751	0.119
7	University of Glasgow	340			0.178	1.818	0.094
8	University of Edinburgh	575			0.375	1.682	0.137
9	Royal Holloway University of London	61			0.004	2.169	0.023
10	University of Warwick	316			0.115	1.831	0.098
11	University of York	334			0.145	1.836	0.091
12	University of Bath	272			0.072	1.854	0.087
13	University of Bristol	272			0.088	1.857	0.088
14	Loughborough University	65			0.003	1.976	0.035
15	University College London	433			0.207	1.761	0.117
16	University of Newcastle	73			0.005	1.992	0.036
17	University of Surrey	214			0.072	1.908	0.067
18	Lancaster University	394			0.177	1.783	0.109
19	University of Durham	176			0.056	1.969	0.057
20	King's College London	240			0.054	1.870	0.077
21	Heriot-Watt University	201			0.069	1.894	0.066
22	Aston University	44			0.005	2.163	0.022
23	University of Westminster	99			0.023	2.063	0.036
24	Massachusetts Institute of Technology	897			1.000	1.506	0.167
25	Stanford University	843			0.838	1.535	0.164
26	University of Maryland	698			0.670	1.615	0.147
27	IEEE	633			0.296	1.650	0.151
28	University of California	546			0.206	1.698	0.138
29	University of Manchester	481			0.231	1.736	0.126
30	IBM	440			0.135	1.757	0.125
31	Microsoft Research	423			0.101	1.768	0.124
32	Carnegie Mellon University	380			0.078	1.791	0.117
33	Harvard University	377			0.134	1.815	0.108
34	University of Nottingham	346			0.181	1.814	0.095
35	University of Liverpool	325			0.108	1.828	0.098
36	Queen Mary University of London	320			0.148	1.824	0.096
37	INRIA	255			0.039	1.860	0.088
38	University of Illinois at Urbana-Champaign	254			0.028	1.867	0.092

Figure 2.3: NodeXL plug-in for Microsoft Excel 2007

Gephi

http://gephi.org/

Gehpi is an interactive network visualisation and exploration tool. Apart from general graph property calculations, its main feature is direct visual graph manipulation. It allows one to interactively select nodes from the visualisation and perform actions. For example, it can select some nodes and combine them into one; change the colour of nodes; draw new nodes to the graph; connect nodes together and rank nodes by colouring them. This software can be very useful when constructing visualisations, but direct graph manipulation requires high-end hardware and is sometimes unstable. It fails to load our test network.

NetworkX

http://networkx.lanl.gov/

This is a Python extension package for network analysis. It enables the Python programming language to load, construct, analyse and visualise networks. Many popular graph analysis functions are provided. This package is a library extension, therefore requires writing Python programs. It is suitable for specific analyses that need to use custom algorithms that are not provided by any other graphical tools.

SNAP

http://snap.stanford.edu/

SNAP is a general purpose network analysis and graph mining library. Like NetworkX, it provides a set of libraries to enable network analysis. It is implemented in C++, therefore would be more efficient than those implemented in a higher level language. Again, it is a programming language extension which requires writing code.

igraph

http://igraph.sourceforge.net/

igraph is yet another programming language extension providing functions that specifically help network analysis. It is programmed in C, with many high level program language interfaces available, including R, Python and Ruby.

2.4.1 Summary

There are two classes of software package in this review. One class is the graphical user interface applications that can load and analyse networks interactively. The best two in this class are the Network Workbench Tool and the NodeXL plug-in for Excel. Both provide standard algorithms to apply to the networks and are capable of visualising networks. Compared to the others, their advantages are the ability to load large networks and perform calculations on those networks in a reasonable time.

The other class is the programming extension libraries that implement standard graph manipulation and analysis algorithms. These include NetworkX, SNAP and igraph. The benefit of doing network analysis at a programming level is that there is no limitation on what one can or cannot do, and it would be no problem to handle large networks. However, the drawback is also clear: it takes time getting started both in terms of learning the language and writing standard analysis functions. It is always beneficial to keep in mind that there is this set of lower-level tools that can become useful when the graphical tools are unable to perform certain desired tasks.

Chapter 3

Graph-Structure Based Author Rankings

3.1 Ranking authors

The PageRank algorithm is an established ranking method used by a successful search engine to rank web pages. The key idea of the PageRank algorithm is instead of ranking pages by the frequency of re-occurring words on a particular web page, it examines link structures of web pages. It ranks a page highly if there are many highly ranked pages linking to it.

The PageRank algorithm can only be applied to directed and unweighed networks. In order to apply it to a undirected graph, Mihalcea[36] suggested a way of converting undirected graph into directed graph by treating each edge as one in-edge and out-edge. Ding [17] applied the PageRank algorithm on the co-citation network to rank authors and they found to be co-cited with an important author boosts the rankings. We apply the algorithm on the co-authorship network in this report.

There are other ways to rank authors by, for example, rank using the h-index[28], the citation count, the collaboration count and betweenness centrality. Since we are trying to rank authors using institutional database, the h-index, the citation count would be incomplete. We used the collaboration count and the betweenness centrality.

Collaboration Count Collaboration count is also called degree centrality. The degree centrality of a node is the number of edges the node has. In a co-authorship network, degree centrality of an author is the size of his collaboration. An author who has co-authored papers with 100 authors would have a bigger collaboration circle than, for example an author who only co-authored with 5 authors. Therefore, the degree centrality ranking can be interpreted as the popularity of an author.

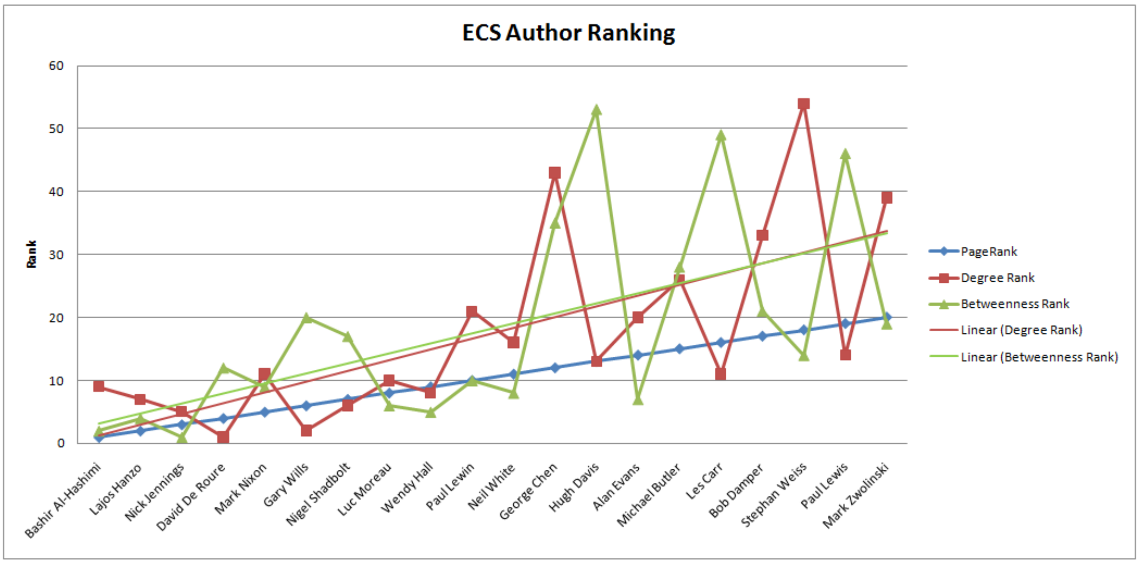


Figure 3.1: The ranking of top ECS authors. Used three algorithms: PageRank, degree centrality and betweenness centrality. The authors are ordered by PageRank.

Betweenness Ranking Betweenness of a node within a network is the number of times this node sits on the shortest path of other two nodes. An author with high betweenness in a co-authorship network means that he is in the centre of information flow for the network. Therefore the betweenness ranking reveals information controller within that network. The higher the author's rank, the more important that author is in passing the information across the network.

We have ranked all the authors who have published papers between 1965 and 2009 within the School of Electronic and computer Science (ECS) in the University of Southampton. Here we compare the three ranking methods – PageRank, degree centrality and betweenness centrality (Figure 3.1).

The chart shows the authors in PageRank order, hence the straight blue line. The red and the green straight lines are the trend lines of the degree and betweenness rankings. All the three ranking methods show an increasing trend. However, degree centrality and betweenness centrality show disagreement for a few authors in their rankings. While Hugh Davis is 12th in degree ranking, he is only 53rd in betweenness ranking; while Stephan Weiss is 14th in betweenness ranking, he is 54th in degree ranking. Interestingly, in these two particular cases, the PageRank gives one vote each – in formal one, PageRank is in agreement with degree ranking, given 12th to Huge Davis; but in latter, PageRank agrees with betweenness ranking, gives 18 to Stephan Weiss.

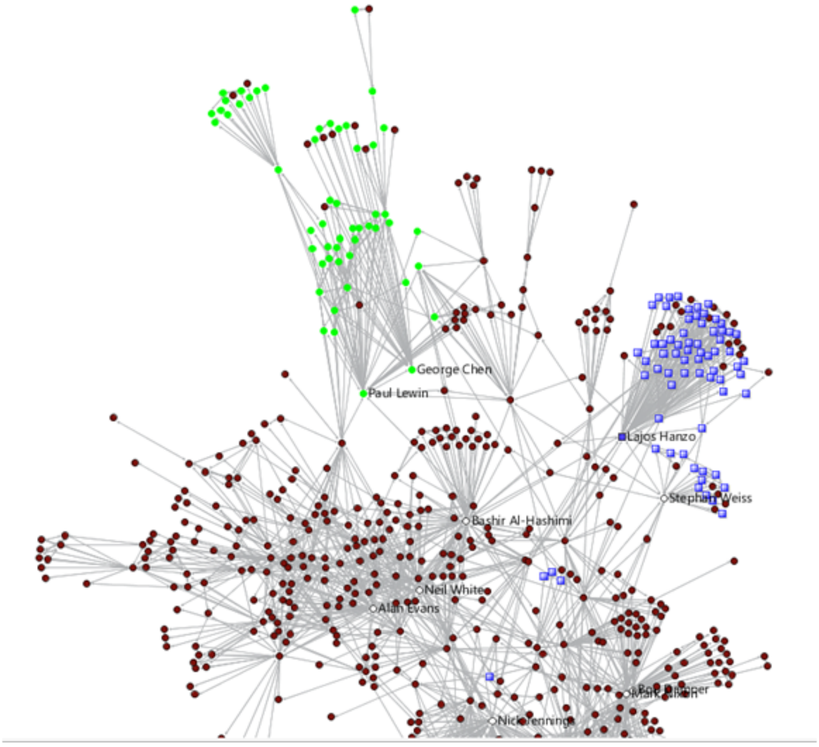


Figure 3.2: Visualisation of ECS co-authorship network. Highly ranked authors are labelled. Two research groups are marked with different node symbol to demonstrate the key positions of these highly collaborative authors in the network

3.1.1 Ranking and Roles

In the top 20 PageRanked authors, 16 of them are professors; 8 (out of 11) heads of research group are presented; 1 Deputy Head of research is also in the list. All these facts show high correlation between the role and the ranking of a researcher. One explanation could be the head of a research group have the opportunity to sign every single paper that the group produces, making them appear to be highly collaborative and hence ranked high.

Figure 3.2 is a visualisation of the collaborations between ECS researchers with top ranked people labelled. On one hand, we could see that these highly PageRanked authors have direct links to almost all members in his group, making them the centre for their group; on the other hand, they hold the 'bridge' position, which they are the only path or shortest path to connect to the outside world, controlling the information flow.

Figure 3.3 shows the heads of ECS research groups arranged at the bottom of the graph, the edges cut off at the top of the graph connects mostly to their group members. Interestingly, although each of them has many connections from their own research group as shown from the dense edges towards the top, they do not have close collaborations. Only three of them have a closed triangle, and one of them is even disconnected from the others.

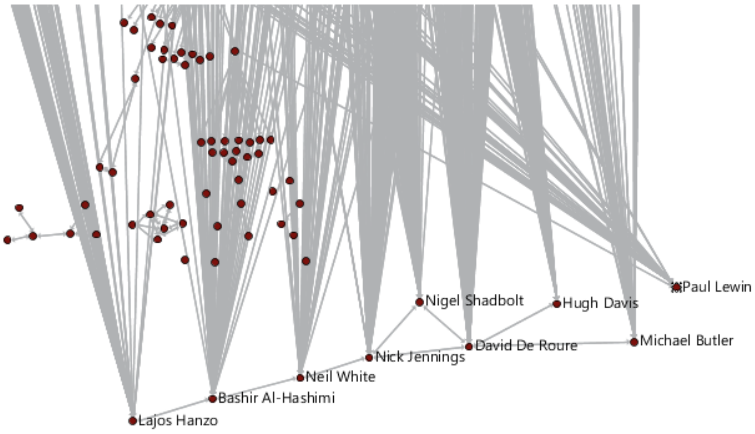


Figure 3.3: Collaborations between Head of research groups is very sparse. All the edges cut off at the top of the graph connects mostly to their group members.

3.2 Discussion

In section 3.1, we described the results of ranking using three different methods – PageRank, degree centrality and betweenness centrality. While three ranking methods show consistency in ordering the authors, they show significant variance at individual levels.

Degree ranking works well on those highly collaborative authors who have many collaborators, but is seriously flawed when it ranks less collaborative author – it is unable to order authors with same number of collaborations. Both PageRank and betweenness rank do not have this problem.

All three rankings are inversely proportional to the number of articles one has published – the more paper an author has, the higher his ranking is. However, it is not the only factor. Lajos Hanzo had published 1003 papers ranked 2 in PageRank, behind Bashir Al-Hashimi who only published 254 papers. If we look at the Lajos's collaboration circle, he has only worked with 72 authors out of his 1003 publications, while Bashir Al-Hashimi has 63 unique co-authors out of 254.

As we have shown in section 3.1.1, in the top 20 authors ranked by PageRank, 8 of them are head of group for the research groups in ECS. This may indicates two situations, first, an ordinary researcher have worked hard, collaborated with many other researchers and published many papers, and as a result, he has got promoted to be the head of the research group.

The other situation could be: The head of group is an ordinary researcher with a similar amount of publications as other members in the group. Since he became the head of the group, many collaboration opportunities arise. As a result, he collaborated more and published more papers. It would be really interesting to carry out analysis to find out which situation applies in each individual case.

Chapter 4

University Collaboration Analysis

As the research projects becoming more open and collaborative, researchers from different universities form groups and produce publications together. If we relate these publications directly to the universities which participated in these collaborations, we could create a university collaboration network.

4.1 General Network attributes

A university collaboration network is created using the publication metadata collected from ACM, which contains 62280 publications between year 1952 and 2010. Figure 4.1 is a visualisation with only the world class universities and top UK universities shown. The edges between the universities are weighted, the thickness represents the amount of collaborations happened between them. This network consists of 1807 universities, institutes and companies. The graph analysis show a tight collaboration between them – the network diameter is 4, the average path length is 2.37 and the entire network is connected with no isolated islands.

From the degree distribution (Figure 4.2), we can see that many affiliations have little collaborations (scattered dots on the right), while most of the collaborations happen in relatively few affiliations. This distribution pattern appears to be follow power-law. However, as shown in figure 4.3, the distribution plotted on a logarithmic scale do now follow a straight line, therefore it is not a power-low distribution. Instead, from the shape of the curve, this distribution shows a bias towards giving more edges to even fewer nodes. What this means is a significant amount of collaborations only happen between very few affiliations; while the remaining affiliations almost don't have collaborations. If we consider the collaborations as the resource or assets available to an affiliation, then this dataset shows that 80% of the resource goes to only 25.7% of the affiliations. Figure 4.4 is the visualisation of the collaboration distribution.

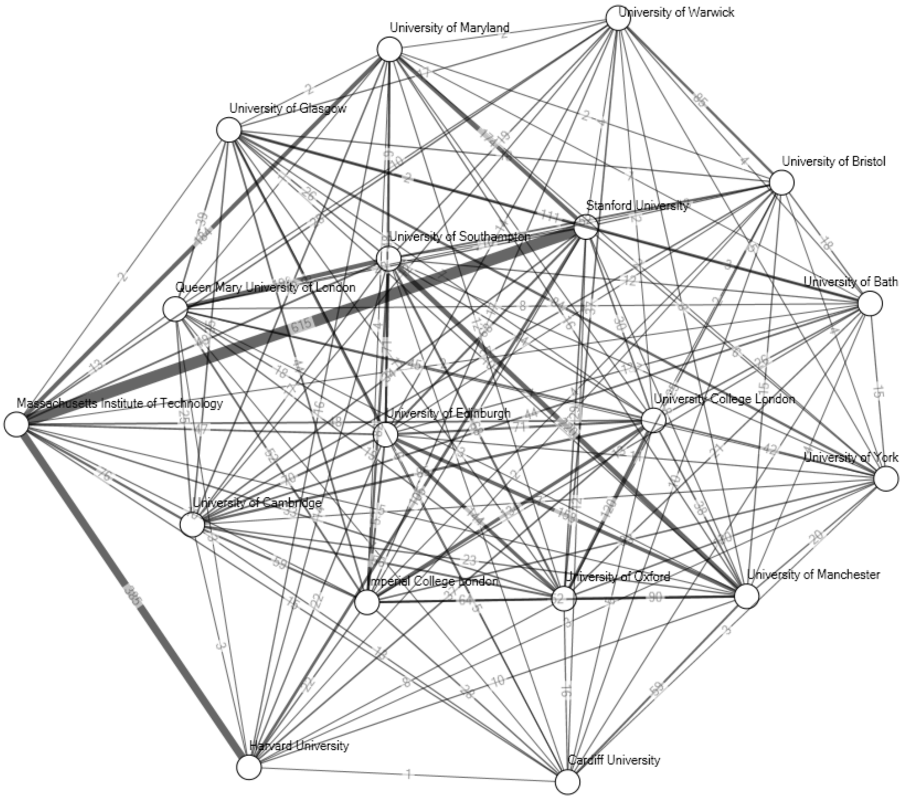


Figure 4.1: University collaboration graph

4.2 University Collaboration Pattern

The previous session shows university collaborations from a quantitative perspective. In this session, we investigate the collaboration patterns only for selected universities.

Due to the data availability, we study the University of Oxford (2010 UK computer science rank 2) and the Cardiff University (2010 UK rank 24) here.

All the affiliations are categorised into 5 groups: The top ten UK universities, the lower ranking UK universities, the top ten world universities (excluding UK), the lower ranking world universities (excluding UK) and non-university. Figure 4.5 and Table 4.1 show the collaboration distribution of these categories for Oxford university and Cardiff university. There are three noticeable differences between the two, first, Oxford has almost equal amount of collaborations with UK top ten and with lower ranking UK universities. This gives significantly more average collaborations with the top ten universities (59 collaborations per affiliation) than with the lower ranking universities (13 collaborations per affiliation) in the UK. This gives evidence that Oxford emphasise collaborations with top UK Universities. On the other hand, Cardiff have 5 times more percentage collaboration with UK lower ranking universities than UK top ten universities, showing no emphasis in

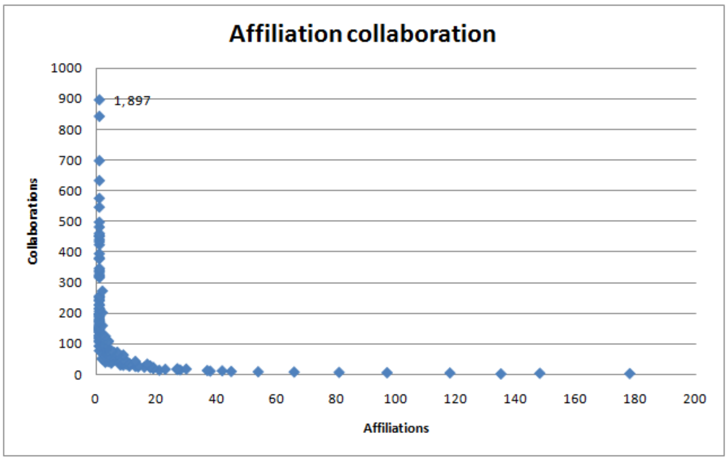


Figure 4.2: Affiliation's collaboration distribution. The most of the affiliations have very few or no collaborations (The dots towards right of the graph), while very few affiliations have most of the collaborations.(The dense dots towards left)

collaborating with high ranking UK universities. On average, Cardiff has collaborated 10 times with top ten UK universities over this period, which is less than with lower ranking UK universities' 14 times.

The second one is the collaboration with world top ten ranking universities. Again, Oxford heavily collaborated with them, producing average collaboration per affiliation of 20; while Cardiff do not show any strong collaboration, with average collaboration just reaching 8 over the same period.

The last difference is that Oxford University has nearly three times more total collaborations than Cardiff University. Within those collaborations, Oxford has twice as many collaborated affiliations than Cardiff.

From these evidences, if we consider Oxford University as a top university, which has many first-class researchers and many funding sources to carry out leading research work; and consider Cardiff University as a lower ranked university with less resources available to it, we can conclude the following:

1. Top universities work closely with top universities – both domestic and world leading universities.

2. Lower ranked university work more frequently with lower ranked universities.

The most interesting observation is that the research quality of a university is not completely defined by itself – how good the researchers are, how well they are equipped – but the universities that it closely works with define it.

These two findings lead to new questions: Do top universities intentionally choose top universities to collaborate with? What are the incentives and constrains that may attach

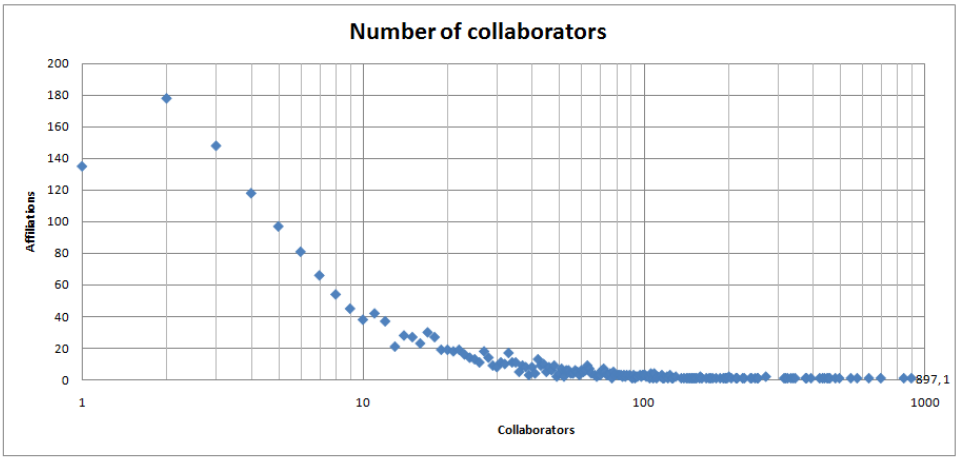


Figure 4.3: Affiliation's collaboration distribution on logarithmic scale. Same data as figure 4.2, but plotted on logarithmic scale and reversed the axis

to individuals within the university when they are choosing who to collaborate with? Are the lower ranked universities willing to collaborate with top universities?

Since similar universities work more closely than others, it should be possible to visualise the split between the universities.

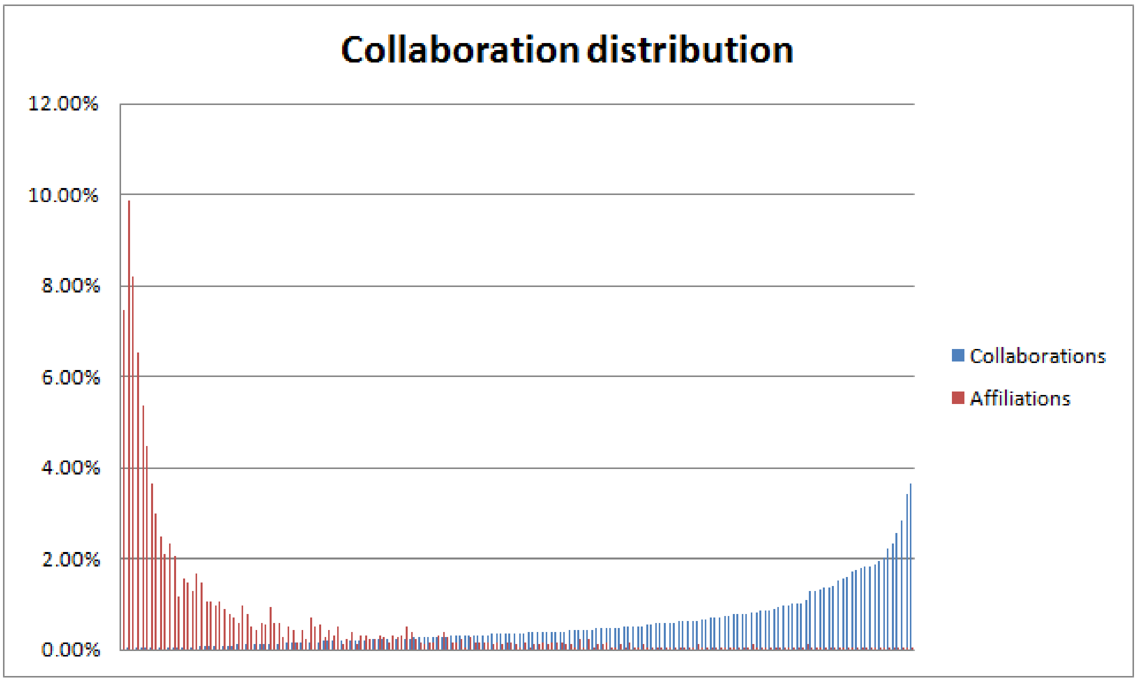


Figure 4.4: Degree distribution among the affiliations. The blue bars represent the number of collaborations, the red bars represents number of universities sharing the amount of collaborations(blue bars) next to the red bars. Towards the right of the diagram, there are high blue bars and low red bars indicating that a lot of collaborations only happen between very few affiliations; towards the left of the diagram, there are high red bars and low blue bars indicating that very few collaborations happen between most of the affiliations.

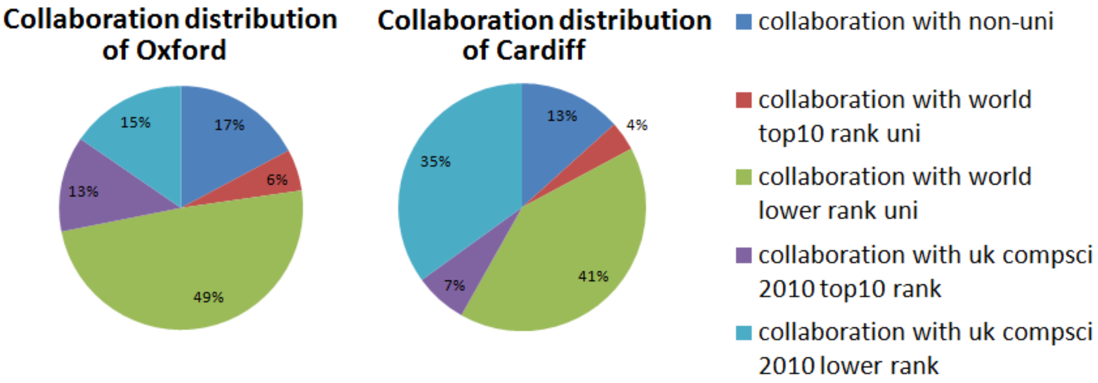


Figure 4.5: Collaboration distribution for University of Oxford (left), and Cardiff University (right)

University of Oxford	Coll.	Coll.Perc.	Affs.	Avg. Coll. per Aff.
Coll. with non-Univ.	556	17%	57	9.75
Coll. with World top ten Univ.	178	6%	9	19.78
Coll. with World lower Univ.	1591	49%	275	5.78
Coll. with UK top ten Univ.	410	13%	7	58.57
Coll. with UK lower Univ.	499	15%	37	13.49
Total	3234	100%	385	8.40
Cardiff University	Coll.	Perc.	Affs.	Avg. Coll. per Aff.
Coll. with non-Univ.	146	13%	21	6.95
Coll. with World top ten Univ.	42	4%	5	8.40
Coll. with World lower Univ.	449	41%	97	4.63
Coll. with UK top ten Univ.	75	7%	7	10.71
Coll. with UK lower Univ.	383	35%	26	14.73
Total	1095	100%	156	7.02

Table 4.1: Collaboration break down for Oxford and Cardiff universities. Collaboration (Coll.) is the number of author level collaboration between the affiliations. Collaboration percentage (Coll.Perc.) is the percentage of this collaboration over total number of collaborations; Affiliations is the number of affiliations it has collaborated with; Average collaboration per affiliation (Avg. Coll. per Aff) collaboration divide by affiliations

Chapter 5

Conclusion & Future Work

This report started by reviewing relevant aspects of social network analysis from different domain perspectives. The mathematicians and physicists' primarily concern is to develop models for networks. They tried to explain observations and properties of various networks from the topological property of the network. Three categories of models were mostly used in studying real life networks – random network models, scale-free network models and small-world network models. The social scientists are more concerned with the relations between people within a social network. The study of the tie strength is an attempt to model the closeness of people's relationship. When social scientists collect data for studies, not only the relational data that connects each person is collected, but also the attribute data that describes the individual is also recorded. With these extra data, they would be able to give more in-depth explanation to an observation.

The second part of the literature review focused on the network studies that were based on the publication data. Publication data is particularly interesting because it is well defined and it can form different types of networks depending on what part of the information was used. The typical three networks which can be constructed using publication data are

- Citation networks
- Co-citation networks
- Co-authorship networks

Citation networks are knowledge networks. Previous citation network studies revealed the structure of knowledge and citing behaviour of researchers. Co-citation networks can be modelled as networks of papers or networks of authors. Analyses enables us to understand the similarities in publications. Co-authorship networks are social network with the assumption that people collaborate to become co-authors. Analyses on these networks reveal the collaboration structure of the scientific community and the publishing patterns across various domains. In addition, co-authorship networks also provide a platform for studying network dynamics.

The last part of the literature review evaluated network analysis tools. The Network Workbench Tool and the NodeXL plug-in were the best because of their ability to handle large networks and they provide a large library of algorithms. Programming extension libraries were also considered and compared in the evaluation. Due to the extra effort involved in writing programs to perform specific analysis, these will only be used if the graphical user interface tool is unable to handle the network.

Chapters 3 and 4 demonstrated two experiments done using publication data. Chapter 3 analysed the author ranking of ECS researchers. Three ranking methods were adopted – PageRank, Collaboration Count Ranking and Betweenness Ranking. The three rankings show consistency in the broad sense, demonstrating correlation between high ranking with high citation, high betweenness and high PageRank. However, they vary significantly at the individual level.

Chapter 4 studied university level collaboration using the publication data. General university collaboration-network analysis shows that the network is tightly connected with a small diameter and small average path length. However, the collaboration distribution is biased towards higher ranking universities which collaborate more with other higher ranking universities. Therefore, we could determine whether a university is highly ranked or not by looking at its composition of collaborators.

Emerging research questions

- As discussed in section 3.2, the head of each research group ranked highly in various ranking methods and were very collaborative. There are two possibilities: one is their role resulted their high ranking; the other is their collaborative research work resulted in them being appointed to the position. It would be really interesting to carry out analysis to find out which situation is applicable for each individual case. One possible method is to collect as much publications as possible for each person, and group their publications into two sets: before head of group and after head of group. It is possible to reveal the answer by analysing and comparing the collaboration metrics of these two sets.

- In section 4.2 we discussed university collaboration patterns. However, to confirm the patterns presented, more university data is needed. The plan is to extract publication data from the ACM website for universities ranked both high and low in 2010 computer science ranking(Table B.1). The complete ranking can be found in appendix A.

- We concluded that universities of similar ranking collaborate more often, therefore it is possible to visualise the clustering between similar universities.

- Based on the assumptions that universities collaborate with one particular class of universities more often than others, we raise a research question: what are the fundamental incentives that a researcher in a university would use to choose a particular researcher in another university that shaped the current university collaboration graph?

- Another direction that my research could develop is to investigate metrics that ranks the conferences or workshops. Unlike journals and universities, for which researchers have developed many ranking methods and metrics, the ranking of conferences or workshops continues to use expert voting. For example, the well recognised and widely used conference ranking from the Australian Research Council(ARC)[2] is performed by deans and experts from the Australian Research and Education Association. The problem is that only the conferences that are known to the voters are ranked fairly, so not many conferences are on the ARC ranking list due to the manual process. Many factors can affect the quality of the conference, an important one that determines the grade of a conference – if voting by experts is given a convincing ranking – is how many experts are attending or have attended. However, to determine experts in any field is a controversial problem, so we propose to use the university ranking as a metric. Therefore, the assumption is that a highly ranked conference should be attended by many highly ranked universities, while lower grade conferences have few delegate from highly ranked universities. The data for past conferences, published papers and author affiliations can be collected from the ACM website. The resulting ranking can be compared and evaluated with the ARC ranking.

Bibliography

[1] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. Network flows: theory, algorithms, and applications, 1993.

[2] ARC. Australian research council, 2010. URL `http://www.arc.gov.au/era/era_journal_list.htm[Accessed2/9/2010]`.

[3] 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`.

[4] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509, 1999.

[5] B. Bollobás. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1):41–52, 1981.

[6] L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008 2008. ISSN 0022-0418. URL `www.emeraldinsight.com/0022-0418.htm`.

[7] T. Brody, L. Carr, Y. Gingras, C. Hajjem, S. Harnad, and A. Swan. Incentivizing the open access research web: Publication-archiving, data-archiving and scientometrics. *CTWatch Quarterly*, 3(3), 2007.

[8] P.J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*. Cambridge University Press, 2005.

[9] AE Cawkell. Citations, obsolescence, enduring articles, and multiple authorships. *Journal of Documentation*, 32(1), 1976.

[10] 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.

[11] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[12] F. Chung and L. Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1(1):91–113, 2004.

[13] J.R. Cole. A short history of the use of citations as a measure of the impact of scientific and scholarly work. *The Web of Knowledge: A Festschri in Honor of Eugene Garfield*, pages 281–300, 2000.

[14] B. Cronin and L. Meho. Using the h-index to rank influential information scientistss. *Journal of the American Society for Information Science and Technology*, 57(9):1275–1278, 2006.

[15] R. De Castro and J.W. Grossman. Famous trails to paul erdős. *The Mathematical Intelligencer*, 21(3):51–53, 1999.

[16] M.A. De Menezes, C. Moukarzel, and TJP Penna. First-order transition in small-world networks. *Arxiv preprint cond-mat/9903426*, 1999.

[17] Ying Ding, Erjia Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–43, November 2009. ISSN 1532-2882.

[18] SN Dorogovtsev, JFF Mendes, and AN Samukhin. Metric structure of random networks. *Nuclear Physics B*, 653(3):307–338, 2003.

[19] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *SIGMOD Rec.*, 34(2):33–40, 2005. ISSN 0163-5808. doi: http://doi.acm.org/10.1145/1083784.1083791.

[20] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

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

[22] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[23] A. Fronczak, P. Fronczak, and J.A. Holyst. Exact solution for average path length in random graphs. *Arxiv preprint cond-mat/0212230*, 2002.

[24] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060): 471–479, 1972.

[25] M. Girvan and MEJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (12):7821, 2002.

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

[27] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E.R. Hilf. The access/impact problem and the green and gold roads to open access. *Serials review*, 30(4):310–314, 2004.

[28] J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569, 2005.

[29] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[30] B. Huberman, D.M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1-5), 2009.

[31] In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97, January 2009. ISSN 0306-4573.

[32] I. King, J. Li, and K.T. Chan. A brief survey of computational approaches in social computing. In *Proceedings of the 2009 international joint conference on Neural Networks*, pages 14–19, 2009.

[33] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462 – 1480, 2005. ISSN 0306-4573. doi: DOI:10.1016/j.ipm.2005.03.012. URL `http://www.sciencedirect.com/science/article/B6VC8-4GCWYR0-1/2/dbf340b78851e32f757fe33c20c5a0b9`. Special Issue on Infometrics.

[34] P. Mählck and O. Persson. Socio-bibliometric mapping of intra-departmental networks. *Scientometrics*, 49(1):81–91, 2000.

[35] Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on facebook, 2009. URL `http://overstated.net/2009/03/09/maintainedrelationships-on-facebook`.

[36] R. Mihalcea, P. Tarau, and E. Figa. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1126. Association for Computational Linguistics, 2004.

[37] M. Molloy and B.A. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2/3):161–180, 1995.

[38] James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213–238, 2004. ISSN 00031224. URL `http://www.jstor.org/stable/3593085`.

[39] C.F. Moukarzel. Spreading and shortest paths in systems with sparse long-range connections. *Physical Review E*, 60(6):6263–6266, 1999.

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

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

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

[43] 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.

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

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

[46] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341 – 346, 1999. ISSN 0375-9601. doi: DOI:10.1016/S0375-9601(99)00757-4. URL `http://www.sciencedirect.com/science/article/B6TVM-3Y6H1T9-N/2/a2cd2aa19f3124e7c79385846149bad0`.

[47] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332–7342, Dec 1999. doi: 10.1103/PhysRevE.60.7332.

[48] Byung-Won On. Social network analysis on name disambiguation and more. In *2008 Third International Conference on Convergence and Hybrid Information Technology (ICCIT)*, number vol.2, pages 1081–8, Los Alamitos, CA, USA, 2008 2008. IEEE Computer Society. ISBN 978-0-7695-3407-7. 2008 Third International Conference on Convergence and Hybrid Information Technology (ICCIT), 11-13 November 2008, Busan, South Korea.

[49] J.P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332, 2007.

[50] M. Ozana. Incipient spanning cluster on small-world networks. *EPL (Europhysics Letters)*, 55:762, 2001.

[51] O. Persson. All author citations versus first author citations. *Scientometrics*, 50(2): 339–344, 2001.

[52] D.J.S. Price. Networks of scientific papers. *Nuovo Cimento*, 5:199, 1957.

[53] D.J.S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[54] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[55] Monica Sharma and Shalini R. Urs. Network dynamics of scholarship: a social network analysis of digital library community. In *Proceeding of the 2nd PhD workshop on Information and knowledge management*, pages 101–104, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-257-3. doi: 10.1145/1458550.1458570. URL `http://portal.acm.org/citation.cfm?id=1458550.1458570`.

[56] R.J. Silverman. Higher education as a maturing field? evidence from referencing practices. *Research in Higher Education*, 23(2):150–183, 1985.

[57] M.E. Soper. Characteristics and use of personal collections. *The Library Quarterly*, 46(4):397–415, 1976.

[58] D. Strauss. On a general class of models for interaction. *SIAM review*, 28(4):513–527, 1986.

[59] Y.M. Su, S.C. Yang, P.Y. Hsu, and W.L. Shiau. Extending co-citation analysis to discover authors with multiple expertise. *Expert Systems with Applications*, 36(3): 4287–4295, 2009.

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

[61] A.F.J. van Raan. Measuring science. capita selecta of current main issues. *Handbook of quantitative science and technology research*, pages 19–50, 2004.

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

[63] H.D. White. Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2):87–108, 2001.

[64] H.D. White and B.C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

[65] 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.

[66] S. Woolgar. Beyond the citation debate: Towards a sociology of measurement technologies and their use in science policy. *Science and Public Policy*, 18(5):319–326, 1991.

[67] Jiadi. Yao. Analysis of university researcher collaboration networks using co-authorship. 2009.

[68] D. Zhao. Going beyond counting first authors in author co-citation analysis. *Proceedings of the American Society for Information Science and Technology*, 42(1), 2005.

Appendix A

List of Universities

This is the list of universities that we aim to collect all their publications that exist in the ACM database. These universities are chosen based on their UK 2010 Computer Science Rankings and the data availability from ACM. Four frequently collaborated world leading universities are also included.

- Aston University
- Cardiff University
- Harvard University
- Heriot-Watt University
- Imperial College London
- King's College London
- Lancaster University
- Massachusetts Institute of Technology
- Queen Mary University of London
- Royal Holloway University of London
- Stanford University
- University College London
- University of Bath
- University of Bristol
- University of Cambridge

- University of Durham
- University of Edinburgh
- University of Glasgow
- University of Liverpool
- University of Manchester
- University of Maryland
- University of Nottingham
- University of Oxford
- University of Southampton
- University of Surrey
- University of Warwick
- University of Westminster
- University of York

Appendix B

UK University Computer Science 2010 Ranking (Times Online)

Ranking	University	Research	Entry Req.	Survey	Prospects	Total
1	Cambridge	4.6	591	82%	93%	100
2	Oxford	4	510		88%	96
3	Imperial College	4.1	467	82%	97%	95.4
4	Southampton	4.1	405	84%	82%	90.8
5	Glasgow	3.8	369	85%	85%	90
6	Edinburgh	4.1	441	78%	86%	89.9
7	St Andrews	2.8	443		89%	88.1
8	Royal Holloway	3.2	292	87%	89%	87.4
9	Warwick	2.9	464	83%	78%	87.1
10	York	3.5	430	76%	89%	86.8
11	Bath	3.5	412	79%	79%	85.5
=11	Bristol	3.6	448	74%	84%	85.5
13	Loughborough	2.6	316	87%	85%	85.3
14	University College London	4	404	75%	81%	85.2
15	Newcastle	3.2	325	84%	81%	85
16	Leeds	3.6	345	79%	80%	84
17	Strathclyde	2.5	381	82%	83%	83.9
18	Aberdeen	3.2	341	78%	86%	83.5
19	Birmingham	3.7	372	79%	73%	83.4
20	Surrey	2.2	350	81%	91%	83.3
21	Sheffield	2.9	367	74%	90%	82.5
22	Manchester	3.9	356	72%	80%	82
23	Dundee	2.9	360	84%	67%	81.9
24	Cardiff	3.2	325	77%	81%	81.4
25	East Anglia	3.1	299	83%	73%	81
26	Lancaster	3.6	319	78%	72%	80.9

27	Bangor	2.5	267		93%	80.6
=27	Durham	3.1	409	67%	90%	80.6
29	Exeter	2.8	318	80%	78%	80.4
30	Nottingham	3.8	310	74%	78%	80.3
31	Essex	2.9	308	81%	72%	79.8
=31	King's College London	2.8	350	75%	81%	79.8
33	Leicester	3.1	278	79%	78%	79.5
34	Heriot-Watt	2.7	347	81%	69%	79.4
35	Kent	2.9	305	75%	82%	78.8
36	Swansea	3.4	309	74%	76%	78.6
37	Liverpool	3.7	321	76%	62%	77.6
38	Sussex	3.2	340	72%	74%	77.4
39	Aberystwyth	3.4	241	80%	67%	77.2
40	Queen's Belfast	2.7	325	71%	82%	77
41	Reading	1.6	332	78%	79%	76.4
42	Hull	1.8	249	81%	80%	75.8
43	Robert Gordon	2	298		81%	75.5
44	Glyndr	1.9			82%	75.1
45	Plymouth	3.5	207	75%	65%	73.7
46	Aston	2	320	74%	72%	73.5
47	Queen Mary, London	3.5	276	67%	69%	72.5
48	Bournemouth	1.8	241	73%	83%	72.3
49	West of England	2.2	252	72%	77%	71.9
=49	Brighton	2.6	246	70%	75%	71.9
=49	Brunel	2.7	302	72%	64%	71.9
52	City	2.6	223	68%	80%	71.1
=52	Ulster	2.4	230	76%	64%	71.1
54	Oxford Brookes	2.5	206	72%	76%	71
55	Stirling	2	261		70%	70.5
56	De Montfort	2.2	187	78%	65%	70.3
57	Greenwich	1.1	169	88%	61%	70.1
58	Portsmouth	1.4	242	74%	71%	68.9
59	Keele		262	81%	73%	68.6
60	Glamorgan	1.8	256	76%	58%	68.3
=60	Teesside	2.4	281	73%	54%	68.3
62	Central Lancashire		213	80%	78%	67.9
=62	Edinburgh Napier	1.4	265		69%	67.9
64	Nottingham Trent	1.5	217	68%	82%	67.7
=64	Glasgow Caledonian	1	300	78%	55%	67.7
66	Salford	2.7	179	72%	63%	67.6
67	Liverpool John Moores	2.2	195	69%	72%	67.4
=67	Hertfordshire	2.5	191	72%	63%	67.4
69	Northumbria		238	81%	68%	67
70	Staffordshire	1.4	207	72%	73%	66.6

71	Coventry	1.6	280		60%	66.4
72	Goldsmiths College	2.9	219	67%	59%	66.2
73	Lincoln	2.5	243	71%	51%	65.7
=73	Manchester Metropolitan	1.7	222	71%	64%	65.7
75	Newman		156		86%	65.3
76	Huddersfield	1.5	246	70%	63%	65.2
77	Kingston	1.7	198	72%	62%	64.9
=77	Cumbria		277	74%	70%	64.9
79	Sheffield Hallam	1.3	223	68%	66%	63.4
=79	Chester		244	73%	70%	63.4
81	Chichester		272		65%	63.1
82	Bedfordshire	1.4	148	74%	60%	63
83	Abertay		314		56%	62
84	Bradford	2	229	63%	61%	61.9
=84	Gloucestershire		179	75%	69%	61.9
86	Middlesex	1.9	144	67%	57%	60.4
=86	Derby		234	68%	70%	60.4
88	Sunderland	1.3	191	69%	55%	59.9
89	Wolverhampton		172	77%	56%	59.7
90	Worcester		178	72%	67%	59.6
91	Anglia Ruskin		240	71%	59%	59.2
92	Northampton		202	71%	61%	58.9
93	London South Bank	1.6	125	76%	38%	58.6
94	Edge Hill		228	67%	63%	58
95	East London		171	71%	62%	57.8
=95	Westminster	1.5	161	68%	50%	57.8
97	UWIC, Cardiff		184		61%	57.4
98	Canterbury Christ Church		150	74%	57%	57.3
99	Southampton Solent		173	70%	61%	57
100	Roehampton		155	73%	55%	56.8
=100	Buckinghamshire New		167	72%	55%	56.8
102	Leeds Metropolitan		207	64%	56%	53.8
=102	Newport		210	62%	58%	53.8
104	Birmingham City		211	67%	47%	53.5
105	Thames Valley	0.9			51%	53

Table B.1: UK University Computer Science 2010 Ranking (Data source: Times Online)