Mathematics Genealogy Networks

University of Oxford

Priya Narayan
Linacre College

Supervisors:
Dr Mason Porter
Dr Elizabeth Leicht

A thesis submitted for the degree of

M.Sc. in Mathematical Modelling and Scientific Computing

2010 - 2011

I, Priya Narayan, hereby declare that the content of this dissertation is entirely my own work, except where otherwise indicated, and all the assistance I have received has been fully acknowledged.

Priya Narayan
357008
Linacre College
September, 2011

"A word to the wise ain't necessary
- it's the stupid ones that need the advice."

"In order to succeed,
your desire for success
should be greater than
your fear of failure."

Bill Cosby

In loving memory of my Dadi G.

Acknowledgements

I would like to take this opportunity to thank both my supervisors, Dr Mason Porter and Dr Elizabeth Leicht for all their undivided support, advice, and guidance throughout this work. I am grateful to them for introducing me to networks. It has been an honour to have had the chance to be supervised by them both. From them, I have learned a great deal through this challenging project, and appreciate their patience with every stumbling block and for being by my side in all complications that I have had. I am grateful to Dr Leicht for also introducing me to coding in Python. Without her advice and tutorials, I would not have been able to construct the data scraper.

I would also like to thank Mitch Keller and the Mathematics Genealogy Project for providing us the data to make this dissertation possible. Without the help of Geoff Evans and his advice on SQL, I may not have been able to extract the Mathematics Genealogy Project data in the first place, so I would like to extend my gratitude towards him.

Dr Kathryn Gillow, the course organiser, has seen me through my tough times and I would like to thank her for her constant moral support and invaluable encouragement throughout the year.

Being in the company of my classmates has been inspirational and motivational, and I would like to thank them for their company through all the good and rough times.

My family have been extremely understanding and would like to thank them for their unfaltering support and encouragement.

Finally, but definitely not least, I am grateful to Professor Marletta, Dr Schmidt, and my other lecturers at Cardiff University for making it possible for me to get here in the first place.

Abstract

Many systems of interest in the physical, biological, and social sciences consist of components linked together and can therefore be modelled as networks. Modelling a system as a network, an abstract structure, captures only the basics of connection patterns and little else. However the connections in a network of people might represent how people learn, form opinions, and gather knowledge, which is of interest in many instances.

The aim of the Mathematics Genealogy Project is to 'compile information about ALL the mathematicians of the world' in the attempt to 'trace the intellectual history of mathematics'[1]. Information on each mathematician is available along with their adviser(s) and advisee(s), which one can use to construct a mathematics genealogy tree.

This mathematics genealogy tree can be modelled as a network, and network theory can be used to identify and gain insights into the patterns in the interactions between mathematicians and their influence on the the structure of the mathematics community.

In this work, a background to networks and concepts used in network theory is given. Using these concepts, the mathematics genealogy tree is modelled by three different networks. Exploring the structure of each network by computing basic diagnostics from network theory, we try understand the influence of advisers on their students.

[1]Taken from the mission statement of the Mathematics Genealogy Project, which is available online (`http://genealogy.math.ndsu.nodak.edu/mission.php`).

Contents

List of Figures

Chapter 1

Introduction

1.1 The Mathematics Genealogy Project

The Mathematics Genealogy Project (MGP) is an online database[1] that contains a wealth of information on individuals who have received doctorates in mathematics. It attempts to trace the intellectual history of mathematics. The aim of the MGP is to list the following information about each individual in the database:

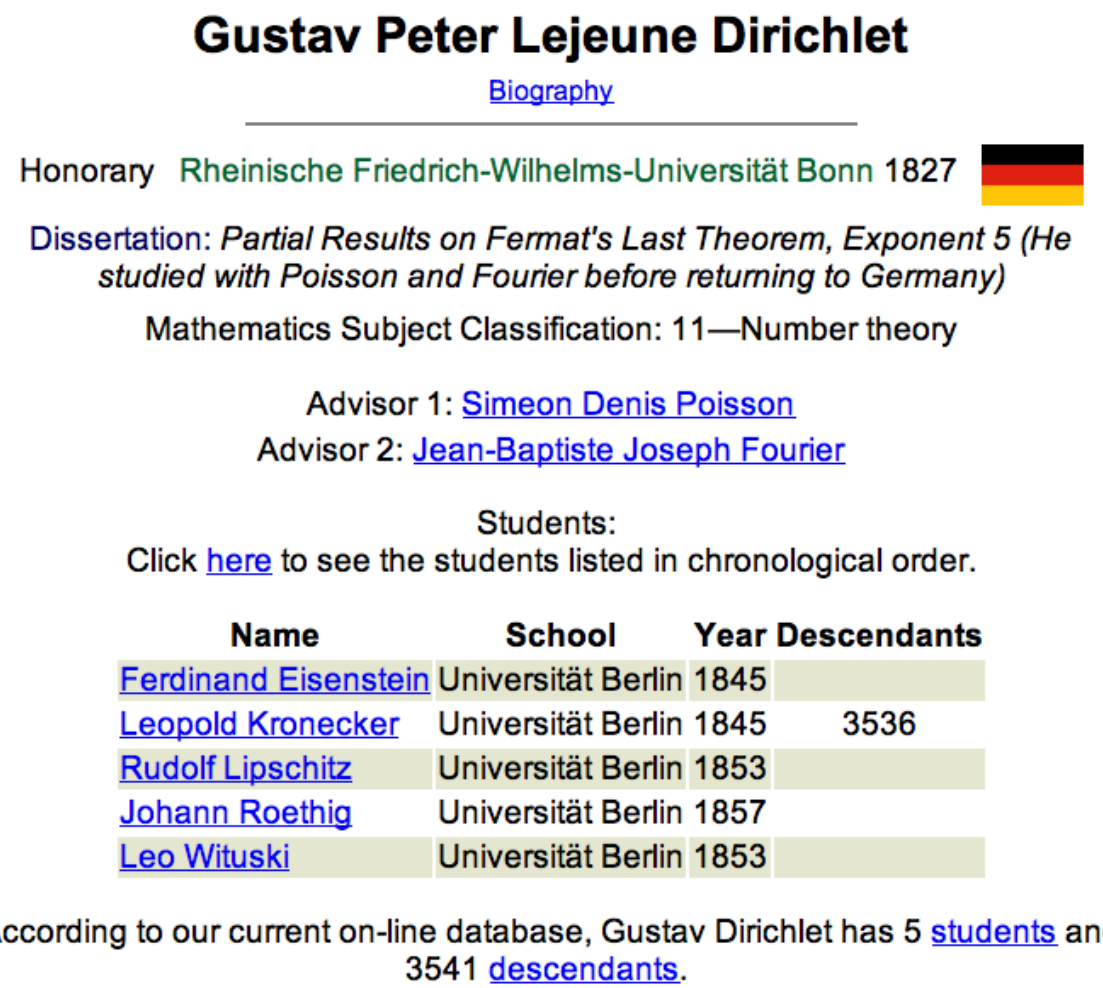
- The **name** of the degree holder
- The **university name** and **country location** that awarded the degree
- The **year** the degree was awarded
- The **title** of their dissertation
- The **Mathematics Subject Classification**[2] (MSC) number of their dissertation
- His/ her **advisor(s)** and their corresponding information
- His/ her **student(s)**[3] and their corresponding information
- The total number of **descendant(s)**.

We refer to the information available in the MGP database as the *MGP data set*. An example of this information represented on the MGP website is given in Figure 1.1, which is a screen shot of Gustav Dirichlet's MGP web page.

[1] `http://genealogy.math.ndsu.nodak.edu/`

[2] This is an alphanumerical classification scheme collaboratively produced by two major mathematical reviewing databases, *Mathematical Reviews* and *Zentralblatt MATH*, which is used by many mathematics journals to classify publications by subject area. See Appendix A for more details.

[3] Also referred to as advisee(s).



Gustav Peter Lejeune Dirichlet

Biography

Honorary Rheinische Friedrich-Wilhelms-Universität Bonn 1827

Dissertation: *Partial Results on Fermat's Last Theorem, Exponent 5 (He studied with Poisson and Fourier before returning to Germany)*

Mathematics Subject Classification: 11—Number theory

Advisor 1: Simeon Denis Poisson
Advisor 2: Jean-Baptiste Joseph Fourier

Students:
Click here to see the students listed in chronological order.

Name	School	Year	Descendants
Ferdinand Eisenstein	Universität Berlin	1845	
Leopold Kronecker	Universität Berlin	1845	3536
Rudolf Lipschitz	Universität Berlin	1853	
Johann Roethig	Universität Berlin	1857	
Leo Wituski	Universität Berlin	1853	

According to our current on-line database, Gustav Dirichlet has 5 students and 3541 descendants.

Figure 1.1: Screen shot of Dirichlet's MGP web page.

1.2 Aim of Dissertation

The MGP data set can be represented as a mathematics genealogy tree, as shown in Figure 1.2[4]. Not only is it interesting to trace back the lineage of a favourite mathematician, but one could also study the structure of this genealogy tree to understand how the mathematical community has developed, how new members join the family of mathematicians, and the influence that advisers have on their advisees. In [1], the MGP data set was used to study the role of mentorship in advisees performance. Also, by examining information such as each individual's university location or mathematics subject area, one can look for patterns within this data and address questions that are of interest. For example, by adding the location from which each individual in the MGP obtained their degree, one could examine the movement of mathematical knowledge around the world, and which location is most popular based on the number of individuals awarded

[4] `http://genealogy.math.ndsu.nodak.edu/`

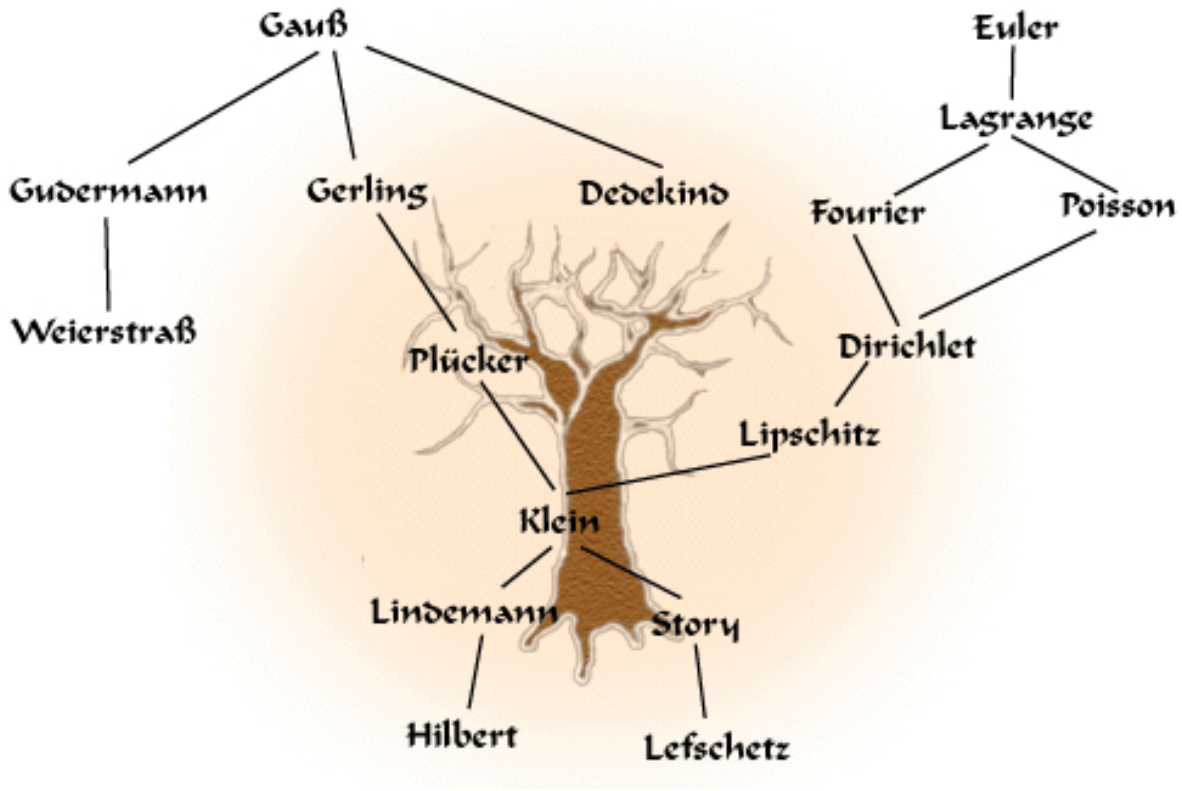


Figure 1.2: Example of a mathematics genealogy tree.

their degree from different countries. Alternatively, one could apply network theory in order to mathematically define the prestige of each university. This is done in [2], for universities in the United States based on the MGP data set.

The aim of this dissertation is to explore the advisor-advisee relationship in the community of mathematicians and infer from this the influence an advisor has on their advisees' supervising behaviour, or how the advisees are influenced by their advisers.

1.3 Proposed Method and Content

The mathematics genealogy tree in Figure 1.2 can be thought of as a network, where the nodes are taken to be individuals in the database, and there exists an edge between two individuals if there is an academic-advising relationship between the two individuals. A variety of useful network diagnostics can be calculated to capture particular features of the network topology and identify behavioural trends between different types of nodes.

The structure of the work presented here is divided into two parts. The first part aims to give a background in the theory of networks used in this work, while the second part represents the MGP data set as a network and applies network diagnostics to explore and better understand the patterns in advisor-advisee relations in the mathematical

community and mathematics genealogy tree.

Part I consists of Chapter 2 and Chapter 3, which are based largely on expositions in Mark Newman's book [6]. In Chapter 2 the various network structures implemented in this work and the methods used to represent and model networks mathematically are discussed. The diagnostics used to help understand particular features and detect characteristics of the networks are given in Chapter 3.

Part II begins with Chapter 4, in which the MGP data set is examined in more detail. There are three different networks considered in this work. Chapter 5 is devoted to the representation of the MGP data set as a directed network, and Chapter 6 for the representation of the data set as two undirected networks. Both Chapters 5 and 6 also include the network diagnostics, given in Chapter 3, computed for each network. All work given in Chapters 5 and 6 is my own, and all the diagnostics were calculated using MATLAB and codes written by me. The results of the diagnostics calculated for each network are summarised in Chapter 7, along with the interpretations in terms of the mathematics genealogy tree and advisor-advisee relations. Chapter 8 discusses and hypothesises possible explanations of the main results mentioned in Chapter 7. Characteristics of an individual in the MGP, other than their number of advisers or number of advisees, has been looked into briefly in Chapter 9. Possible scope for further work is also suggested in this chapter.

The original dissertation proposal is given in Appendix A, which also includes details of all progress that was made and the reason for changing focus. Appendix B briefly states the statistical knowledge of discrete distributions used in this work. A summary of only the basic diagnostics calculated throughout this work is given in a table in Appendix C.

Part I

An Introduction to Network Theory

Many systems of interest in the physical, biological, and social sciences that consist of individual parts or components linked together in some way can be represented in the form of a network. A network is essentially a collection of points connected together in pairs. For example, the Internet is a collection of computers that are linked by data connections. In human societies, the people could be thought of as nodes in the network, connected by acquaintances or some form of social interaction. The pattern of the connections between the nodes and the structure of networks can tell a lot about the behaviour of the system that it represents.

The connections in a social network affect how people learn, form opinions, and gather knowledge. Unless something is known about the structure of a network, we cannot hope to fully understand the functions of the corresponding systems.

http://thecustomizewindows.com/2011/03/communication-approaches-on-social-networks/

Chapter 2

Structure and Mathematical Representation of Networks

2.1 Networks and their Representation

A *network*, also referred to as a *graph* in mathematical literature, is a collection of items called *nodes*, with connections between them called *edges*. The number of nodes in a network is commonly denoted as n, and m is used to denote the number of edges. For the small network given in Figure 2.1, the total number of nodes n is 7 and the number of edges m is 6.

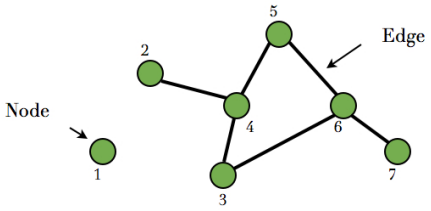


Figure 2.1: An example network with 7 nodes and 6 edges.

There are several ways to represent a network mathematically. But first, the nodes need to be labelled uniquely with integer labels $1, \ldots, n$ (see Figure 2.1) to be able to refer to or select a node in the network without any ambiguity. The order in which the nodes are labelled is not important. The *set of nodes* is denoted by $\mathcal{N}$ here. An edge between nodes i and j can be denoted by (i, j). Using this notation, the complete network can be specified by giving the value of n and a list of all of the edges, called an *edge list*. For example, the small network given in Figure 2.1 has $n = 7$ nodes and can be represented mathematically by the following edge list:

$$\{(2,4),(3,4),(3,6),(4,5),(5,6),(6,7)\}.$$

An element in the edge list, i.e. a pair of nodes (i, j) can be assigned a unique numerical label from $1, \ldots, m$. We call the set of these numerical labels for the edges of a network the *set of edges*, which we shall denote as $\mathcal{E}$ here. Edge lists are useful to store the structure of a network, as they are compact. However, to compute network diagnostics, the *adjacency matrix* is a better representation of a network, as many network diagnostics are tied to concepts from linear algebra. The definition of the adjacency matrix of a network depends on how an edge is defined in the network (see Section 2.2.2).

2.2 Network Structure

2.2.1 Simple Graphs and Multigraphs

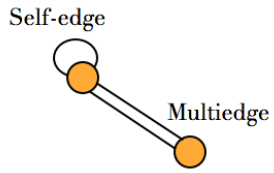


Figure 2.2: A self-edge and a multiedge.

Edges that connect nodes to themselves are called *self-edges*. Self-edges are not considered in this work, as it does not make sense for individuals to advise themselves to obtain a doctorate in mathematics. If there is more than one edge between the same pair of nodes, then the edges are called a *multiedge*. A network with multiedges is called a *multigraph*. Figure 2.2 gives an example of a self-edge and a multiedge. A network that has neither self-edges nor multiedges is called a *simple network* or a *simple graph*.

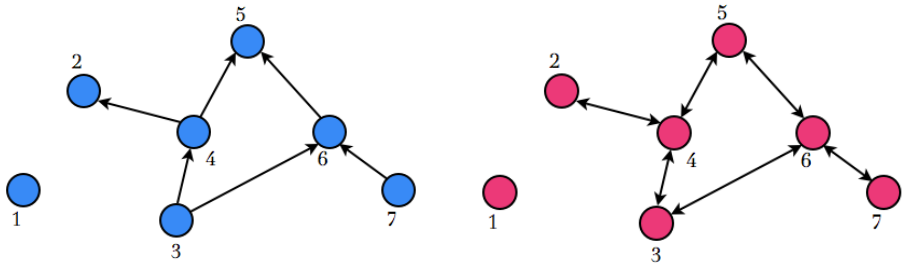
2.2.2 Directed and Undirected Networks

In some networks, like Figure 2.3a, edges can have a direction, that points from one node, the 'source', to another, the 'target' node. For example, node 2 in Figure 2.3a would be the target node and node 4 the corresponding source node associated to the edge between nodes 2 and 4. Such networks are called *directed networks* or *directed graphs* or *diagraphs* for short, and the edges are called *directed edges*. *Undirected networks* can be thought of as directed networks in which each undirected edge has been replaced with two directed ones running in opposite directions between the same pair of nodes (as shown in Figure 2.3b).

An element of the adjacency matrix of a directed network is given by

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge } \textit{from} \text{ node } j \textit{ to} \text{ node } i, \\ 0 & \text{otherwise.} \end{cases}$$

By convention, the direction of the edge runs from the second index to the first index [6]. Hence, the adjacency matrix for the directed network in Figure 2.3a is



(a) Directed Network (b) Undirected Network

Figure 2.3: Example of a directed network and its undirected counterpart.

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \textit{Nodes} & \textit{1} & \textit{2} & \textit{3} & \textit{4} & \textit{5} & \textit{6} & \textit{7} \end{matrix} \\ \begin{matrix} \textit{1} \\ \textit{2} \\ \textit{3} \\ \textit{4} \\ \textit{5} \\ \textit{6} \\ \textit{7} \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{2.1}$$

An element of the adjacency matrix for an undirected network is given by

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the adjacency matrix for the network in Figure 2.3b is

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \textit{Nodes} & \textit{1} & \textit{2} & \textit{3} & \textit{4} & \textit{5} & \textit{6} & \textit{7} \end{matrix} \\ \begin{matrix} \textit{1} \\ \textit{2} \\ \textit{3} \\ \textit{4} \\ \textit{5} \\ \textit{6} \\ \textit{7} \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}. \tag{2.2}$$

There are a few points to note about the structure of the adjacency matrix for both directed and undirected networks. The diagonal entries of an adjacency matrix are always zero if a network has no self-edges. The adjacency matrix for an undirected network is always symmetric. In fact, the adjacency matrix for an undirected matrix can be constructed from its directed counterpart by making the adjacency matrix for the directed

network symmetric. The number of edges, m, in an undirected network is the same as that in its directed counterpart. However, for an undirected network, each edge is counted twice in the adjacency matrix, as $A_{ij} = 1$ if there is an edge between i and j. Therefore, the total number of edges for an undirected network is given by

$$m = \frac{1}{2} \sum_{ij} A_{ij}. \tag{2.3}$$

For example, the sum of the elements of the adjacency matrix (2.2) for the undirected network given in Figure 2.3b is 12, which is twice the number of edges $m = 6$. In a directed network with no self-edges, each edge is counted once in the adjacency matrix, so the number of edges in the directed network is given by

$$m = \sum_{ij} A_{ij}. \tag{2.4}$$

Hence, for the directed network given in Figure 2.3a, from the diagram it can be seen that there are 6 edges, which is also the sum of the elements of its adjacency matrix given in (2.1).

2.2.3 Directed Acyclic Graphs

A *cycle* in a directed graph is a closed loop of edges with arrows on each of the edges pointing the same way around the loop, as shown in Figure 2.4.

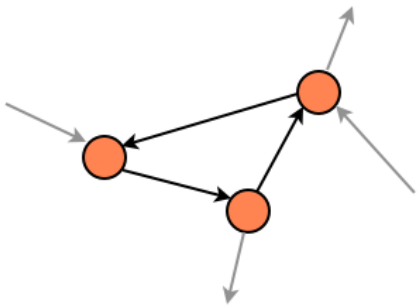


Figure 2.4: An example of a cycle in a directed network.

A *directed acyclic graph (DAG)* is a directed network in which there are no cycles. A family tree is an example of a directed acyclic graph as, unless one has a time machine, it is impossible for someone to be a biological child and a biological grandparent to the same person or to go back and start the family lineage.

Chapter 3

Network Diagnostics

If the type of edges in the network is known, we can calculate a variety of useful quantities that capture particular features of the network.

3.1 Degree Diagnostics

The concept of *centrality* is used to understand which nodes in the network are the most important [6]. There are many possible ways to define importance, but the simplest centrality measure in a network is to look at the number of edges connected to each node and is referred to as *degree centrality*.

3.1.1 Undirected Network

The *degree* of a node in an undirected network is the number of edges connected to it.

Node Degree The degree for a node i, in an undirected network of n nodes, is denoted by k_i and can be written in terms of the adjacency matrix as

$$k_i = \sum_{j=1}^{n} A_{ij}, \tag{3.1}$$

i.e. the sum of the i^{th} row of the adjacency matrix. Because the adjacency matrix, $\mathbf{A}$, of an undirected matrix is symmetric, this is also the same as taking the sum of the column, i.e.

$$k_i = \sum_{j=1}^{n} A_{ji}.$$

For example the degree of node 4 given by Figure 2.3b can be counted from the diagram to get $k_4 = 3$. Summing the 4^{th} row or column of the corresponding adjacency matrix of the network, given by (2.2), also gives $k_4 = 3$.

Mean Degree The mean degree of a node in an undirected network is

$$c = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

This expression for the mean degree can be simplified and written in terms of the total number of edges in the undirected network. Using (3.1), we can rewrite the double sum in (2.3) so that

$$m = \frac{1}{2} \sum_{i=1}^{n} k_i,$$

which yields

$$c = \frac{2m}{n}. \tag{3.2}$$

Degree Distribution The distribution of the degree of nodes is one of the most basic of network properties. The fraction of nodes in the network that have degree k is denoted by p_k and is given by

$$p_k = \frac{\text{number of nodes with degree } k}{n}. \tag{3.3}$$

The set of these quantities, $\{p_k\}$, gives the degree distribution, and it can be insightful to plot the degree distribution of a large network as a function of k.

That said, the degree distribution does not tell us the complete structure of a network. For example, the two networks in Figure 3.1 have the same degree distribution but are different.

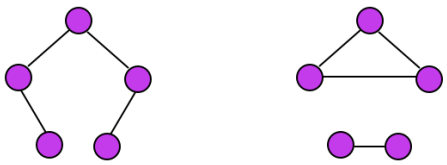


Figure 3.1: Different networks with the same degree distribution.

3.1.2 Directed Network

In a directed network, a node is associated with two types of degree, the *in-degree* and the *out-degree*.

Node Degrees The *in-degree* of a node is the number of incoming edges connected to a node, and because an element of the adjacency matrix of a directed matrix, A_{ij}, is 1 if there is an edge from j to i, the in-degree can be written as

$$k_i^{\text{in}} = \sum_{i=1}^{n} A_{ij}. \tag{3.4}$$

The *out-degree* of a node is the number of outgoing edges from a node and similarly can be written as

$$k_j^{\text{out}} = \sum_{j=1}^{n} A_{ij}. \tag{3.5}$$

Note the change in the summation index from (3.4) to (3.5), which implies that the in-degree of the i^{th} node is the i^{th} row sum of the adjacency matrix and the out-degree of the i^{th} node is the i^{th} column sum of the adjacency matrix.

Mean Degree The mean in-degree and the mean out-degree are given by

$$c_{\text{in}} = \frac{1}{n}\sum_{i=1}^{n} k_i^{\text{in}} \quad \text{and} \quad c_{\text{out}} = \frac{1}{n}\sum_{j=1}^{n} k_j^{\text{out}}, \tag{3.6}$$

respectively. However by substituting (3.4), (3.5), and (2.4) in the above, it can be seen that $c_{\text{in}} = c_{\text{out}}$ and that the mean degree can be written as

$$c = \frac{m}{n}, \tag{3.7}$$

where $c = c_{\text{in}} = c_{\text{out}}$.

Degree Distribution As there are two different degrees associated with each node in a directed network, there are also two different degree distributions in a directed network: the in-degree and out-degree distributions. The method to construct these is the same as discussed for undirected networks. The in-degree distribution is represented by the set $\{p_{k^{\text{in}}}\}$, where $p_{k^{\text{in}}}$ is the proportion of nodes with in-degree k^{in}. Similarly, the out-degree distribution shows the spread of out-degrees of nodes in the network.

The true degree distribution of a directed network could be thought of as the joint distribution of in- and out- degrees [6], which shall be discussed in the degree assortativity in a directed network.

3.2 Assortativity

Another central concept in the study of networks is the correlation between the properties of the nodes connected directly by a single edge, (i.e. nodes that are nearest neighbours). In social sciences, *homophily* designates the tendency of people to associate with others whom they perceive as being similar to themselves in some way [6]. A network shows *assortative mixing* if there is tendency of similar nodes to be connected to each other. *Disassortative mixing* is the tendency for nodes to associate with others who are unlike themselves. The structural properties of a network can be effected profoundly by assortative mixing [4]. For example, in social networks, the patterns of friendship are strongly affected by language and age among other factors. It has been observed that people have a high tendency to have friendship connections with those who speak the same language as themselves [6].

In [4] and [6], assortative mixing has been categorised into 2 groups, 'assortative mixing of discrete characteristics' and 'assortative mixing by scalar properties'. In the first of the 2 groups listed, discrete characteristics can be classified using any alphanumeric labelling scheme, and in the second group, scalar properties can be both discrete or continuous. However, because the discrete characteristics in the first group can be classified by enumeration, assortative mixing by discrete characteristics becomes a special case of assortative mixing by scalar properties. However the diagnostics described for each of the types of assortative mixing is different, and for this reason, the naming convention for each of these groups are kept the same as in [6] to distinguish between the two groups. The assortative mixing discussed below follows that given in [4].

3.2.1 Assortative Mixing of Discrete Characteristics

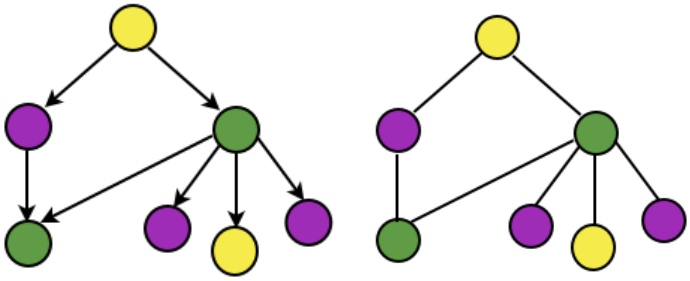
In a network, if the nodes are classified according to some discrete set of characteristics that are enumerative (i.e. they do not fall in any particular order), for example geographical location, then assortative mixing can be quantified by an *assortativity coefficient*, which can be defined in terms of a *mixing matrix*.

In [5], an element of the mixing matrix, E_{ij}, is defined as the number of edges that connect nodes of types i and j. The mixing matrix is symmetric on an undirected network, and it can be asymmetric on directed networks. In this work, we shall explicitly define the mixing matrix for directed networks by

$$E_{ij} = \text{number of edges that connect source nodes of type } j \text{ to target nodes of type } i.$$

The interpretation of this definition of the mixing matrix for both a directed and an undirected network is illustrated by an example. Consider the directed network and its

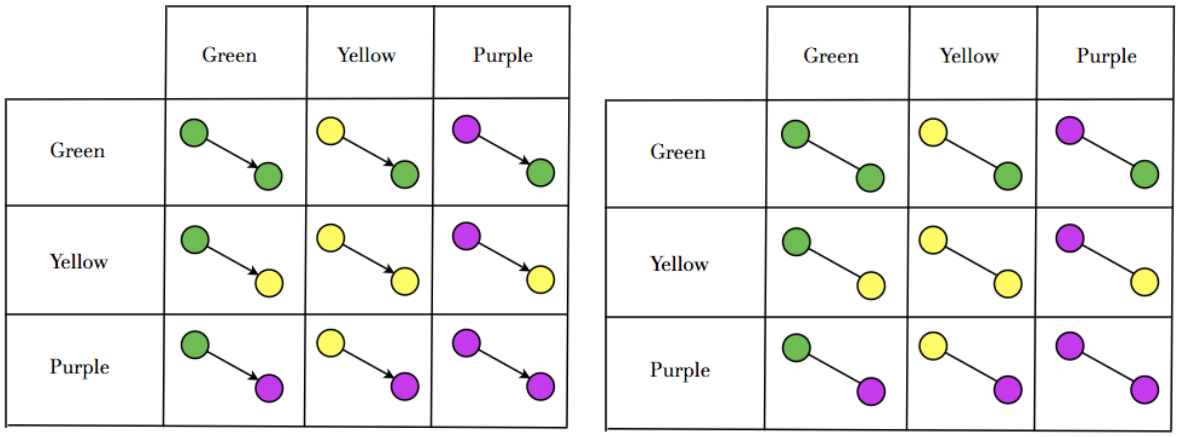
undirected counterpart, given in Figure 3.2, in which there are three types of mixing characteristics distinguished by the colour (green •, yellow •, or purple •) of the node.



(a) Directed Network (b) Undirected Network

Figure 3.2: Example networks in which there are three types of mixing characteristics distinguished by the colour of the node.

The combination of node types that should be counted for each element of the mixing matrix for both the directed network and the undirected network is indicated in Figure 3.3.



(a) Mixing matrix for the directed network (b) Mixing matrix for the undirected network

Figure 3.3: An example: The type of edge and node combination that should summed for each element in the mixing matrix **E**.

Hence, the mixing matrix for the directed network is given by

$$\mathbf{E} = \begin{matrix} \textit{Node Colour} \\ G \\ Y \\ P \end{matrix} \begin{matrix} \begin{matrix} G & Y & P \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 2 & 1 & 0 \end{pmatrix} \end{matrix},$$

and the mixing matrix for the undirected network is given by

$$\mathbf{E} = \begin{array}{c} \begin{array}{cccc} \textit{Node Colour} & G & Y & P \end{array} \\ \begin{array}{c} G \\ Y \\ P \end{array} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix} \end{array},$$

where G represents a green node, Y a yellow node, and P a purple node.

In an undirected network, the edges have no direction, so $E_{ij} = E_{ji}$ and the mixing matrix is symmetric. However, for a directed network, E_{ij} may not necessarily be equal to E_{ji}. The normalised mixing matrix measures the fraction of edges that connect nodes of different types and is given by

$$\mathbf{e} = \frac{\mathbf{E}}{\|\mathbf{E}\|}, \tag{3.8}$$

where the matrix norm $\|\cdot\|$ used is taken as the sum of all of the elements of the matrix [5]. The normalised mixing matrix $\mathbf{e}$ can be thought of as a joint distribution of node types i and node type j, because its elements satisfy

$$\sum_{ij} e_{ij} = 1.$$

Using the mixing matrix, we can define the probability distributions of the types of nodes at the ends of an edge by

$$a_i = \sum_j e_{ij} \quad \text{and} \quad b_j = \sum_i e_{ij}. \tag{3.9}$$

In an undirected network, $a_i = b_i$. However, for a directed network, $\{a_i\}$ can be interpreted as the probability distribution of the type of the target node and $\{b_i\}$ as the probability distribution of the type of the source node.

The assortativity coefficient given in [4], lies in the range $[-1, 1]$ and is defined as

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} = \frac{\text{Tr } \mathbf{e} - \|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|}, \tag{3.10}$$

as $\sum_i e_{ii} = \text{Tr } \mathbf{e}$ and

$$\sum_i a_i b_i = \sum_i \left[\left(\sum_j e_{ij} \right) \left(\sum_k e_{ki} \right) \right] = \sum_{jk} \sum_i e_{ki} e_{ij} = \sum_{jk} \mathbf{e}^2 = \|\mathbf{e}^2\|.$$

An assortativity coefficient value of 0 corresponds to no assortative mixing, as this happens when $e_{ij} = a_i b_j$, implying $\sum_i e_{ii} = \sum_i a_i b_i$ and resulting in the numerator of (3.10) being equal to zero. There is perfect assortative mixing if $r = 1$; this happens when

$\sum_i e_{ii} = 1$. If the network is perfectly dissortative, then r is negative and (according to [4]) takes the value

$$r_{\min} = -\frac{\sum_i a_i b_i}{1 - \sum_i a_i b_i} = -\frac{\|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|}, \tag{3.11}$$

because Tr $\mathbf{e} = 0$ (no like-for-like mixing), and the diagonal of the matrix $\mathbf{e}$ indicates the proportion of edges that join similar nodes.

3.2.2 Assortative Mixing by Scalar Properties

Assortative mixing can also be done according to scalar properties (e.g. age) of a network node. Analogously to Section 3.2.1, we can define a normalised quantity e_{ij} as the fraction of edges that connect nodes associated with a value of j to a node of value i. The values that i and j take could be either discrete (making e_{ij} elements of a matrix, just as described in Section 3.2.1) or continuous, in which case e_{ij} is a function of two continuous variables. The concepts used for the discrete case can be generalised to the continuous case, but in this work we shall only consider the discrete case as given in [4].

In the discrete case, the matrix e_{ij} can be used to calculate the standard *Pearson correlation coefficient*, a measure of assortativity defined by

$$r = \frac{\sum_{ij} ij(e_{ij} - a_i b_j)}{\sigma_a \sigma_b}, \tag{3.12}$$

where, σ_a and σ_b are the respective standard deviations[1] of $\{a_i\}$ and $\{b_j\}$, the probability distributions of the edges that end and start at nodes with values i and j, given by (3.9). Similar to the assortativity coefficient defined in Section 3.2.1, the Pearson correlation coefficient given in (3.12) also lies in the range $[-1, 1]$, where a value of 1 indicates perfect assortative mixing and a value of -1 indicates perfect disassortativity.

3.2.3 Degree Assortativity

A special case of assortative mixing by a scalar node property is mixing by node degree and is referred to as degree correlations in [5]. With this type of assortativity, we can see if nodes of high degree preferentially associate themselves to other nodes of high or low degree. Mixing by node degree can be quantified using the Pearson correlation coefficient given by (3.12).

[1] Appendix B gives the formulas used to determine the standard deviations of a discrete probability distribution.

Undirected Networks

For undirected networks, the matrix used to calculate the Pearson correlation coefficient, for degree assortativity, [3] has entries

$$e_{xy} = \text{the proportion of edges that connect nodes of degrees } x \text{ and } y, \tag{3.13}$$

which shall be referred to as the *degree distribution matrix*[2]. Because e_{xy} is a symmetric matrix, the associated probability distributions (the corresponding a_i and b_j given in (3.12)) are the same. Hence we denote the associated probability distributions of the degrees of nodes as $\{q_x\}$, where

$$q_x = \sum_y e_{xy}. \tag{3.14}$$

The assortativity coefficient for mixing by node degree in an undirected network is, therefore, given by

$$r = \frac{\sum_{xy} xy(e_{xy} - q_x q_y)}{\sigma_q^2}, \tag{3.15}$$

where σ_q^2 is the variance of the distribution $\{q_x\}$.

Thinking of $\{q_x\}$ in terms of a network, it is in fact the distribution of one less than the node degree, also called the *excess degree distribution*. It can be written in terms of the degree distribution $\{p_x\}$, given in (3.3), by

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j j p_j}. \tag{3.16}$$

In [3], the Pearson correlation coefficient is rewritten in terms of the degrees of the nodes at the ends of edges. If the degrees of the nodes at the ends of the i^{th} edge of an undirected network are denoted by x_i and y_i, then the Pearson correlation coefficient for an undirected network with m edges can be given by

$$r = \frac{\frac{1}{m}\sum_{i=1}^m x_i y_i - \left[\frac{1}{m}\sum_{i=1}^m \frac{1}{2}(x_i + y_i)\right]^2}{\frac{1}{m}\sum_{i=1}^m \frac{1}{2}(x_i^2 + y_i^2) - \left[\frac{1}{m}\sum_{i=1}^m \frac{1}{2}(x_i + y_i)\right]^2}. \tag{3.17}$$

Note that the summations in (3.17) are taken over the edges of the network.

Directed Networks

In a directed network, the mixing by node degrees becomes more complex, as each node has both in- and out-degrees. There are at least 4 different ways to define the Pearson

[2]The notation convention for the indices have changed from i, j to x, y in subsection only for undirected networks. This is done in order to avoid confusion in Section 6.4, where quantities stated in this particular subsection are rewritten in terms of the adjacency matrix.

correlation coefficient, by considering the different combinations of the degrees taken at the ends of a directed edge. Mark Newman [4] defined the assortativity coefficient for degree correlations in directed networks as

$$r = \frac{1}{\sigma_q^{\text{in}} \sigma_q^{\text{out}}} \left[\sum_{jk} jk(e_{jk} - q^{\text{in}}(j) q^{\text{out}}(k)) \right], \tag{3.18}$$

where an element of the degree distribution matrix e_{jk} is defined as the proportion of directed edges with a source node of out-degree k and target node of in-degree j. In (3.18), $q^{\text{in}}(k)$ is the proportion of directed edges with a target node of in-degree k and $q^{\text{out}}(k)$ is the proportion of directed edges with a source node of out-degree k. Also, σ_q^{in} and σ_q^{out} are the standard deviations of $\left\{q^{\text{in}}(k)\right\}$ and $\{q^{\text{out}}(k)\}$, respectively. Note that by the definition of the expectation of a discrete distribution, given in Appendix B, (3.18) can be rewritten as

$$r = \frac{1}{\sigma_q^{\text{in}} \sigma_q^{\text{out}}} \left[\sum_{jk} jk e_{jk}^{\text{out}} - \mu_q^{\text{in}} \mu_q^{\text{out}} \right], \tag{3.19}$$

where μ_q^{in} and μ_q^{out} are the expectations of the distributions $\left\{q^{\text{in}}(k)\right\}$ and $\{q^{\text{out}}(k)\}$, respectively. The Pearson correlation coefficient (3.19) measures the tendency of nodes to connect to other nodes that have a similar out-degree to their in-degree. It can sometimes be more useful to consider an assortativity that measures the tendency of nodes connecting to other nodes with similar out-degrees to their own out-degree. This is called *out-assortativity*. Also, *in-assortativity* refers to the tendency of nodes connecting to other nodes with similar in-degrees to themselves. The following out- and in-assortativity coefficients are taken from [7].

Out-Assortativity Constructing the *out-degree distribution matrix*, $\mathbf{e}^{\text{out}}$, with its entries, e_{jk}^{out} taken as the proportion of directed edges with a source node with an out-degree of k and a target node with an out-degree of j, we can define the probability distribution $\{q^{\text{out}}(k)\}$ of a directed edge with a source node that has an out-degree of k as

$$q^{\text{out}}(k) = \sum_{k} e_{jk}^{\text{out}},$$

and the probability distribution $\{q'^{\text{out}}(k)\}$ of a directed edge with a target node that has an out-degree of k as

$$q'^{\text{out}}(k) = \sum_{j} e_{jk}^{\text{out}}. \tag{3.20}$$

The *out-assortativity coefficient* is then defined as

$$r_{\text{out}} = \frac{1}{\sigma_q^{\text{out}} \sigma_{q'}^{\text{out}}} \left[\sum_{jk} jk e_{jk}^{\text{out}} - \mu_q^{\text{out}} \mu_{q'}^{\text{out}} \right], \tag{3.21}$$

where σ_q^{out} and $\sigma_{q'}^{\text{out}}$ are the standard deviations of $\{q^{\text{out}}\}$ and $\{q'^{\text{out}}\}$ respectively, and μ_q^{out} and $\mu_{q'}^{\text{out}}$ are the expectations of the distributions $\{q^{\text{out}}\}$ and $\{q'^{\text{out}}\}$ respectively.

In-Assortativity Similarly, we can construct an *in-degree distribution matrix*, $\mathbf{e}^{\text{in}}$, with the entries e_{jk}^{in} taken as the probability of a directed edge with a source node that has an in-degree of k and a target node that has an in-degree of j. The probability distribution $\left\{q^{\text{in}}(j)\right\}$ of a directed edge with a target node that has an in-degree of j is then given by

$$q^{\text{in}}(k) = \sum_j e_{jk}^{\text{in}},$$

and $\left\{q'^{\text{in}}(j)\right\}$ is the probability distribution of a directed edge with a source node that has an in-degree of j. It is given by

$$q'^{\text{in}}(k) = \sum_k e_{jk}^{\text{in}}.$$

The *in-assortativity coefficient* can then be defined as

$$r_{\text{in}} = \frac{1}{\sigma_q^{\text{in}}\sigma_{q'}^{\text{in}}}\left[\sum_{jk} jk e_{jk}^{\text{in}} - \mu_q^{\text{in}}\mu_{q'}^{\text{in}}\right], \tag{3.22}$$

where σ_q^{in} and $\sigma_{q'}^{\text{in}}$ are the standard deviations of $\left\{q^{\text{in}}\right\}$ and $\left\{q'^{\text{in}}\right\}$ respectively, and μ_q^{in} and $\mu_{q'}^{\text{in}}$ are the expectations of the distributions $\left\{q^{\text{in}}\right\}$ and $\left\{q'^{\text{in}}\right\}$ respectively.

3.3 Clustering

Clustering is an important property in social networks. It is often found that if a node i is connected node k and node k is connected to node j, then there tends to be a high probability that node i is connected to node j [6]. A path of length two consists of three nodes and two edges and is constructed as shown in Figure 3.4 by the solid edges. The

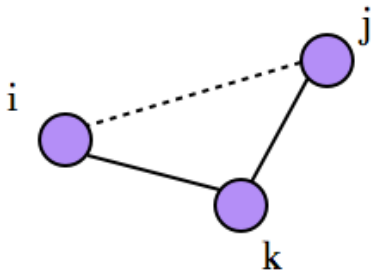


Figure 3.4: A path of length two (solid edges) is closed if the dashed edge is present.

path is *closed* if it forms a triangle, as shown in Figure 3.4 if the dashed edge exists and

is called a *loop* of length three. Transitivity is a special type of clustering and can be quantified by the *clustering coefficient*, which is defined as the fraction of paths of length two in the network that are closed:

$$C = \frac{\text{number of loops of length 3 in the network}}{\text{number of paths of length 2 in the network}}. \tag{3.23}$$

A clustering coefficient can also be defined locally for each node in the network. For node i, the *local clustering coefficient* in [6] is given by

$$C_i = \frac{\text{number of loops of length 3 in which } i \text{ participates}}{\text{number of paths of length 2 for which } i \text{ is the central node}}. \tag{3.24}$$

For nodes with degree 0 or 1, the numerator and denominator are zero; in these cases, the local clustering coefficient C_i is 0. The clustering coefficients lie in the range $[0, 1]$, where a coefficient of 0 implies no transitivity in the network and 1 implies complete transitivity.

Part II

Mathematics Genealogy Networks

Chapter 4

Description of the Data Set

The data set used in this work has been extracted from an SQL database provided[1], of the data underlying the Mathematics Genealogy Project website. This data set consists of 137,138 individuals who have acquired a doctorate in mathematics from 1363 up to 2012.[2] A total of 138,167 advisor-advisee relations are listed. An individual in the data set can have up to a maximum of 5 advisers and up to a maximum of 103 advisees. However, these extreme cases are rare in the data set (see Section 5.2).

Issues with the data

For each individual in the data set, there are fields that indicate the year, the name of the university, the country, and the subject area in which they were awarded their degree. However some of these fields are empty for some individuals. Table 4.1 shows the actual amount of information available. The first column in Table 4.1 lists the different characteristic information for each individual. The second column indicates the total number of different categories of a characteristic. For example there are 61 different countries listed as the location from which an individual was awarded their degree. The third column is the number of individuals in the data set that have non-empty fields. For example, 46,369 of the 137,138 individuals, that is 34% include the subject classification of their dissertation (MSC), and the other 66% have empty fields.

Although not all of the 137,138 individuals have information on all the four characteristics listed in Table 4.1, all individuals are included when constructing the three networks. The fields that are missing for each individual are filled in with zeros, so that in any diagnostic calculated involving any of the associated characteristics of the individuals, the numerical label of zero for the characteristic, represents the group of the

[1]The data has been kindly provided to us by Mathematics Genealogy Project and Mitch Keller.

[2]In the data set, there are 4 individuals with 2010 listed as the year they were awarded their degree, no individuals with 2011 and only 1 individual with 2012.

Characteristic	Number of categories	Individuals that have associated characteristic
Year	463	92% (125,708)
University	659	20% (27,154)
Country	61	22% (29,623)
Subject Area (MSC)	97	34% (46,369)

Table 4.1: Information available in data set.

individuals with no information.

Year degree awarded

Only 125,708 out of the 137,138 individuals in the data set have information about the year they were awarded their degree. Of these, 376 individuals have two or more years associated to them, which are not necessarily consecutive years. Upon inspection, the few individuals that were checked have multiple degrees awarded to them. Based on this finding, for these 376 individuals, the earliest year is taken for the purpose of subsequent computation, as it indicates the first of the multiple degrees the individual was awarded.

4.1 Method of Labelling Nodes

The 137,138 individuals in the data set can be identified by a unique numerical identifier that lies in the range 1 to 139,228. Due to the structure of the database, the individuals are labelled according to this unique numerical identifier. This implies that the adjacency matrix for any network where the nodes are taken to be the individuals will be of a size 139,228 by 139,228. The dimensions of the adjacency matrix will therefore be larger than the number of nodes, 137,138, and will contain columns and rows of zeroes. A numerical scheme is also implemented to classify the different types of a characteristic. For example, each country listed in the database is assigned a unique number label.

4.2 Basic Trends over Time

We can use the year an individual was awarded their degree to explore how diagnostics that involve calculating quantities for each individual, changes over time. However, since not all the individuals in the data set have a year associated to them, it is useful to understand how the individuals are grouped over different time periods. In this section we look at the number of individuals awarded their degree in different time periods, and the number of advisers an individual had for their degree over time. The data

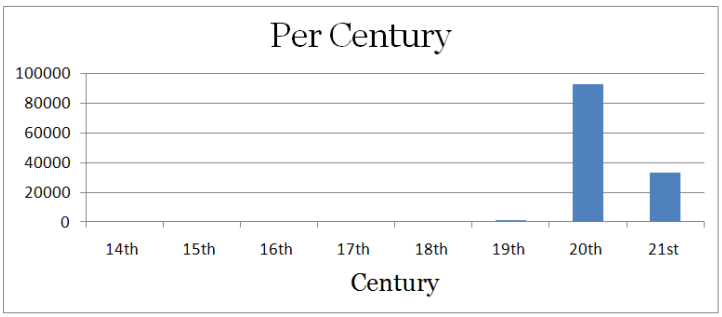
presented in this section is based only on the 125,708 individuals who have information listed on the year they were awarded their degree. Therefore caution must be taken when interpreting the results here, as the remaining 8% of the individuals in the data set could have been awarded their degree any time, and adding their details, if it were available, could influence the results.

Figure 4.1 helps to understand when individuals were awarded their degree and hence how the mathematics genealogy tree grows over time. Let $\mathfrak{C}_i$ denote the number of individuals that were awarded their degree in the i^{th} century. Figure 4.1a is a plot of $\mathfrak{C}_i$ for $i = 14, \ldots, 21$. It seems that the majority of individuals in the MGP data set were awarded their degree in the 20^{th} century based on Figure 4.1a. This result would stand even if the 11,430 individuals that are missing information on their year, were to be associated with any century, because the next largest $\mathfrak{C}_i$ is $\mathfrak{C}_{21}$ which takes the value of around 30,000 (60,000 less than $\mathfrak{C}_i$). Hence adding 11,430 more individuals would not change the fact that the majority of the individuals in the MGP data set were awarded their degrees in the 1900s.

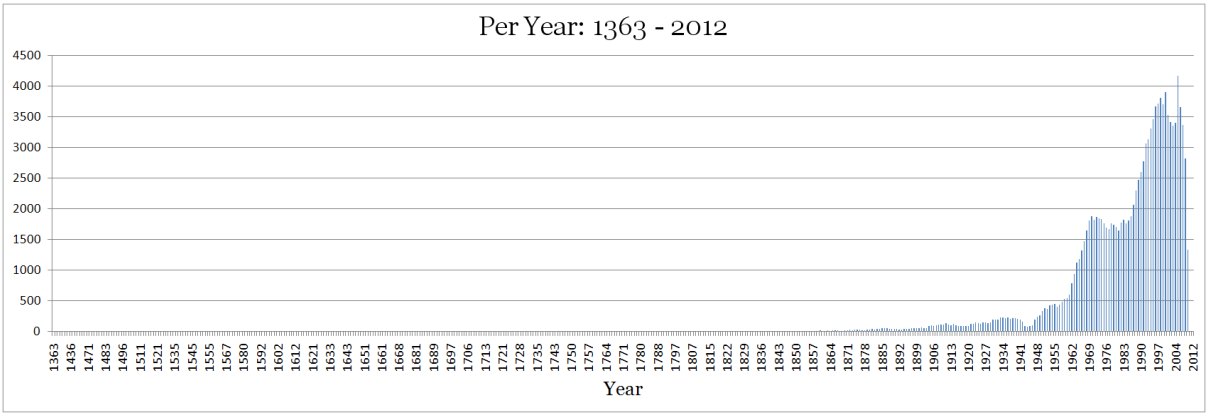
The number of individuals that were awarded their degree in year j has been plotted in Figure 4.1b for $j = 1363, \ldots, 2012$. It is also evident from Figure 4.1b that only a small number of individuals were awarded their degrees prior to the 20^{th} century. Hence when looking at network diagnostics over time, it is important to bear in mind that statistically significant trends cannot be drawn from diagnostics that are calculated pre 20^{th} century for such small number of individuals. And so caution must be taken when looking at the early periods pre 20^{th} century. However Figure 4.1b further indicates that the number of degrees awarded per year in the MGP data set began to grow from the late 1800s, with a significant increase in the mid 1900s to the 1970s. The number of degrees awarded decreases slightly per year in the 1970s, but shortly after, it continues to increase rapidly up until 2009. The numbers fall rapidly in the 2010s indicating that the data set for this period is incomplete.

Figure 4.1c is a truncated and enlarged version of Figure 4.1b that plots the number of individuals per year from 1860 onwards, in which some of the trends observed in Figure 4.1b can be seen in more detail. There is a steady gradual increase in the number of degrees awarded per year from 1860 to 1959, but with a dip over the post-World War II years from 1944 to 1947. From 1959 onwards the number of degrees awarded per year is greater than 500. From Figure 4.1c one can see that over the short period of 1959 to 1970, the number of degrees awarded per year increased from a little over 500 to approximately 1800. After which the number of individuals awarded a degree per year remained in the range of 1500 to 2000 up until the year 1984. In the period 1984 to 1998, the number grew per year to 3800. After which the numbers began to decrease gradually over a period to

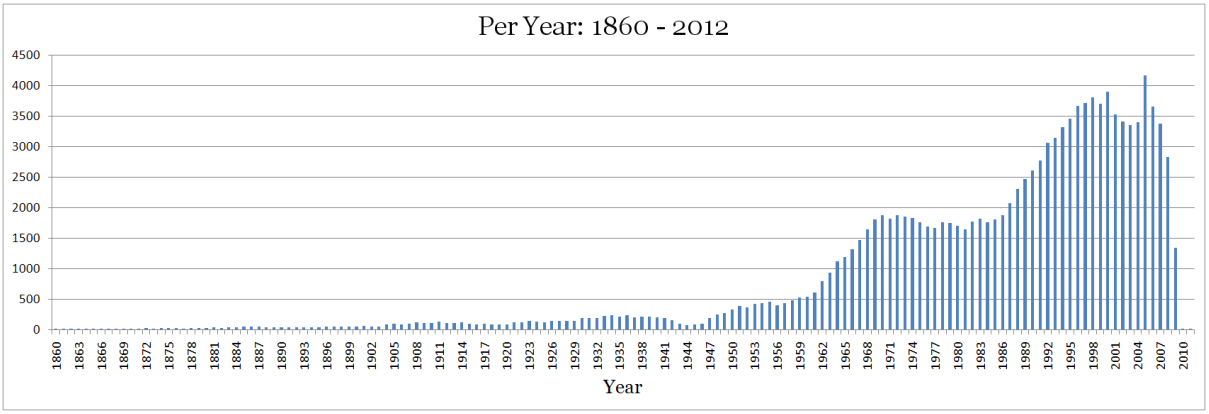
a value of 3400 observed in 2004, with an anomaly in year 2000 when it peaked. After seeing the largest number of individuals awarded in a year in 2005, the numbers decrease rapidly from 2006 to 2012. One could speculate that this recent decrease from 2006 to 2010 could be due to the recent downturn in the economy.



(a) For each century.



(b) For years 1363 to 2012.



(c) For years 1860 to 2012.

Figure 4.1: Number of individuals awarded a degree over time.

Figure 4.2 shows how the number of advisers an individual has had in the data set has changed over each century. For each century, the number of individuals are grouped by the number of advisers they have had, and the proportion of individuals in each of

these groups is plotted in Figure 4.2. If we denote $\mathfrak{C}_j$ for the set of individuals that were

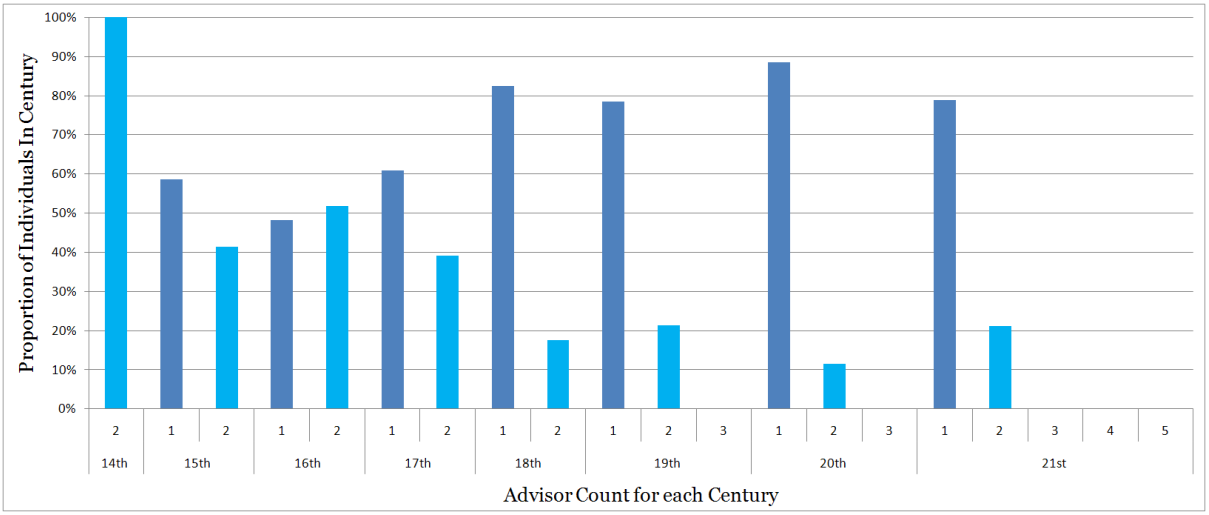


Figure 4.2: Number of advisers an individual has over time (proportion of individuals).

awarded their degree in the j^{th} century, the proportion of individuals in century j with i advisers is calculated as

$$\frac{\text{Number of nodes with } i \text{ advisers}}{|\mathfrak{C}_j|}.$$

Figure 4.2 indicates that the maximum number of advisers an individual has had, has increased in the past 3 centuries. (The proportion of individuals with more than two advisers are so small that they are not visible in Figure 4.2 in the 19^{th}, 20^{th} and 21^{st} centuries.) Although it is interesting to compare the proportions for the earlier centuries, one has to bear in mind that these figures are based on small number of individuals in that time period. We consider the data with statistical significance to begin in the late 1800s if not from the 20^{th} century onwards. In each of the 19^{th}, 20^{th}, and 21^{st} centuries, the majority of individuals ($\sim 80\%$) had just 1 advisor.

Chapter 5

Mathematics Genealogy as a Directed Network

The most natural formulation of the mathematics genealogy tree (see Figure 1.2 as an example of one) is as a directed network, where the nodes are the individuals and the direction in the edges represents the direction of advice, from which one can infer the transfer of information from advisor to advisee.

5.1 Adjacency Matrix

We can encapsulate the mathematics genealogy data as a directed network, in which there is a directed edge from every advisor to each of their advisees. Hence the elements of the adjacency matrix for this directed network are given by

$$A_{ij} = \begin{cases} 1 & \text{if individual } j \text{ advised individual } i, \\ 0 & \text{otherwise.} \end{cases}$$

This directed network is also a directed acyclic graph, by the nature of advising an individual. An advisor can never be a descendant of their student, so cycles cannot form. The clustering coefficient defined in Section 3.3, is thus always zero for a directed acyclic graph. For this reason, the clustering coefficient has not been considered for this network representation of the MGP data set. The diagonal entries of the adjacency matrix are zero, as there are no self-edges due to the fact that one cannot supervise oneself to be awarded a degree.

5.2 Degree

The mean degree of this network is $c \approx 1.0075$. Figure 5.1 is a plot of the in-degree (number of advisers an individual has) distribution of this directed network.

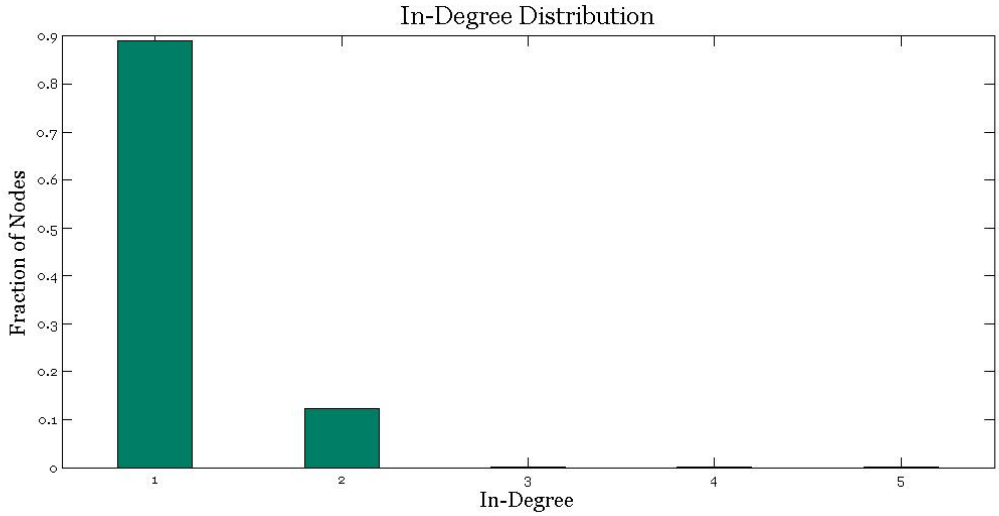


Figure 5.1: In-degree distribution of the directed network.

Figure 5.2 is a plot of the out-degree (number of advisees an individual has) distribution of this directed network, which is also plotted against a log-log scale in the top right corner of the same figure.

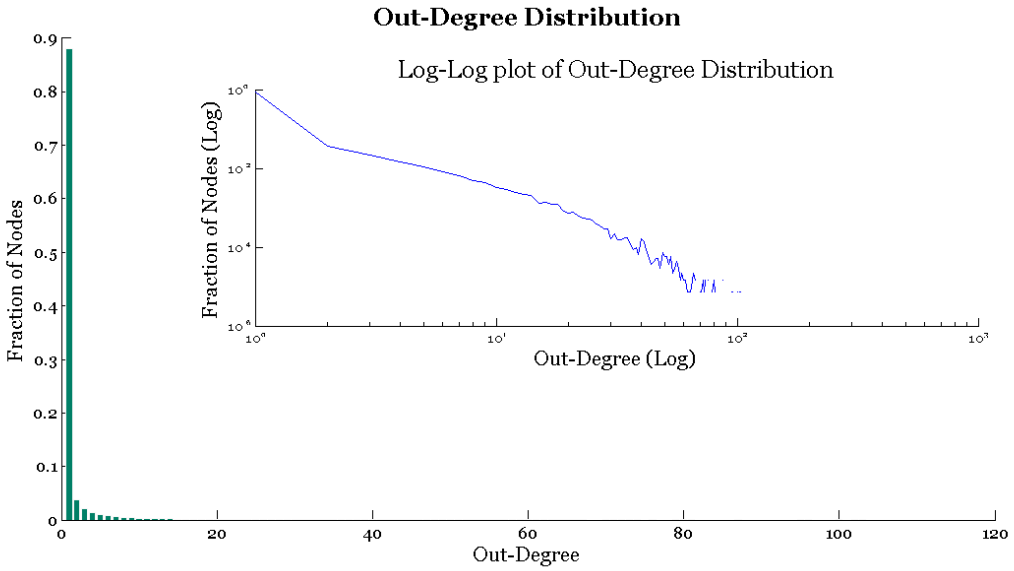


Figure 5.2: Out-degree distribution of the directed network.

The in-degree and out-degree distributions of the directed network given in Figures 5.1 and 5.2 show that most of the nodes in the network have low in- and out-degrees, as expected by the mean degree of $c \approx 1.0075$. From Figure 5.2, one can see that the out-degree distribution has a significant 'tail' to the distribution, corresponding to nodes with substantially higher out-degrees. Comparing the highest in-degree to the highest

out-degree, indicates that in the mathematics community, it is possible for individuals to have many more advisees (103) than advisors (5).

5.3 Out- and In-Degree Assortativity

From the adjacency matrix of this directed network, $\mathbf{A}$, the elements of the degree distribution matrices are computed as

$$e_{jk}^{\text{out}} = \frac{\text{Number of directed edges from a node with out-degree } (k-1) \text{ to a node with out-degree } (j-1)}{\text{Total number of directed edges in network}},$$

$$\text{for} \quad j,k = 1,2,\ldots,(\text{maximum out-degree } +1),$$

$$e_{jk}^{\text{in}} = \frac{\text{Number of directed edges from a node with in-degree } (k-1) \text{ to a node with in-degree } (j-1)}{\text{Total number of directed edges in network}},$$

$$\text{for} \quad j,k = 1,2,\ldots,(\text{maximum in-degree } +1).$$

Figure 5.3 is a visual representation of $\mathbf{e}^{\text{out}}$, the 104×104 out-degree distribution matrix[1], where maroon(■) indicates a value of zero, and the colour gradation green (■) to blue (■) represents values increasing from 0 to 1.

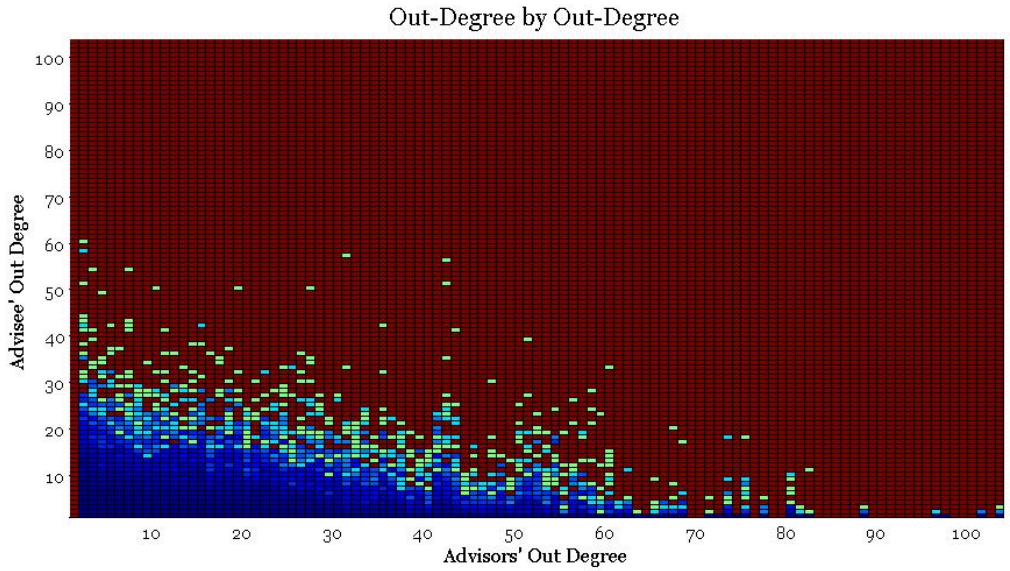


Figure 5.3: Visual representation of the out-degree distribution matrix (order of rows reversed).

[1]Note that the y-axis in Figure 5.3 increases from bottom up, so this visual representation is actually of $\mathbf{e}^{\text{out}}$ but the rows are in reverse order.

The maximum in-degree of the directed network is 5, so the in-degree distribution matrix is a 6×6 matrix because some individuals have an in-degree of zero. The in-degree distribution matrix is given explicitly by

$$\mathbf{e}^{\text{in}} = \frac{1}{138167} \begin{array}{c} \\ \textit{0} \\ \textit{1} \\ \textit{2} \\ \textit{3} \\ \textit{4} \\ \textit{5} \end{array} \begin{array}{c} \begin{array}{cccccc} \textit{In-degree} & \textit{0} & \textit{1} & \textit{2} & \textit{3} & \textit{4} & \textit{5} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 11,558 & 80,430 & 11,948 & 28 & 0 & 0 \\ 6,712 & 20,827 & 6,472 & 10 & 1 & 0 \\ 53 & 89 & 26 & 0 & 0 & 0 \\ 3 & 3 & 2 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{5.1}$$

Since assortativity considers connected nodes, the first column of $\mathbf{e}^{\text{out}}$ represents source nodes (advisers) with an out-degree of zero, and since there are no such edges, the first column is zero. Similarly, the first row of $\mathbf{e}^{\text{in}}$ represents target nodes (advisees) with an in-degree of zero. However, by the definition of a target node (advisee), a target node always has at least one in-degree. Hence the first row entries of $\mathbf{e}^{\text{in}}$ are all zero.

The out-assortativity coefficient (3.21) is calculated to be 0.2188 and the in-assortativity coefficient (3.22) is calculated to be 0.8373, for this directed network. Both Figure 5.3 and the out-degree assortativity coefficient $r_{\text{out}} \approx 0.2188$ only slightly greater than zero, indicate a slight assortative mixing by out-degrees in the directed network. In terms of the mathematics genealogy tree and advisor-advisee relations, advisees who have many academic siblings[2] have a tendency to go on to become advisers with many of their own students. Also, advisers who advise a small number of students have a slight tendency to influence their advisees to have small number of students themselves.

The fact that $r_{\text{in}} \approx 0.8373$, which is not only positive but close to 1, indicates a strong assortative mixing by in-degree, so advisers with a high in-degree have a tendency to associate with advisees with high in-degree. In terms of advisor-advisee relations, this means that advisers who were supervised by many individuals (advisers with high in-degree) have a tendency to advise a student with other individuals (advisees with high in-degree). That said, the in-assortativity coefficient might be a bit of an over statement of the results, as r_{in} is really dominated by the $e^{\text{out}}_{1,1}$ element of the in-degree distribution matrix.

Evolution of r_{out} and r_{in} Over Time

Here we consider how the out- and in-degree assortativity coefficients change over time. The year associated to a node refers to the year the individual was awarded their degree.

[2]Academic siblings are individuals who have the same adviser.

The assortativity coefficients are calculated using edges with advisees associated with the earliest year (t_0) to a final time T that varies from t_0 to the latest year in the data set. Therefore, as T increases, more individuals are added to the network for which each r_{out} and r_{in} are calculated. In effect, the degree assortativity coefficients are cumulative figures, calculated as more advisees are added to the network over time.

The year of the advisee is taken and not of the adviser to define the inclusion of edges in the network, for which the assortativity coefficients are calculated. This is because we want to examine the advising relationship (influence of advisers on the supervising behaviour of their advisees) and it makes more sense to include the advisees in the network.

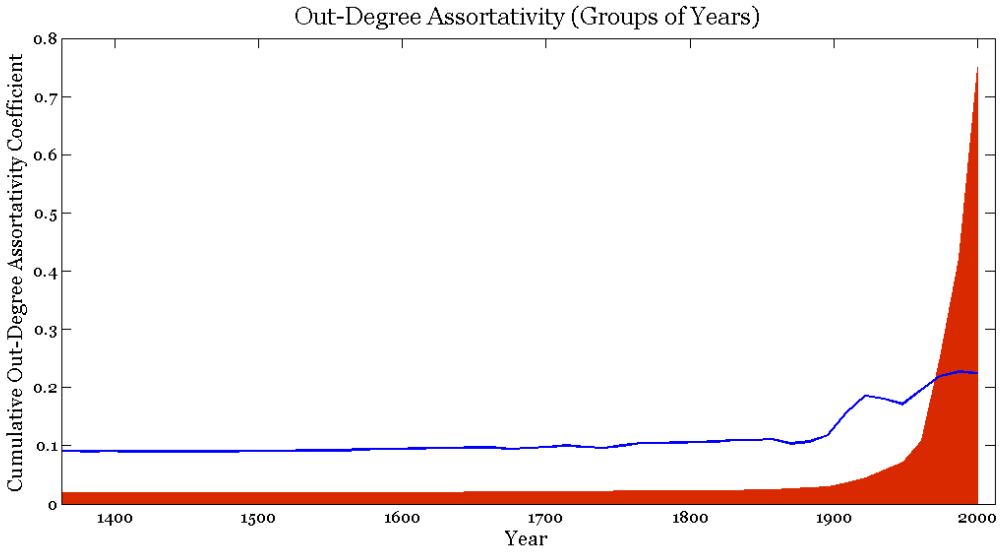
In Figures 5.4 and 5.5, the red area-plot indicates the proportion of edges in the entire directed network included in the calculation of the assortativity coefficients to give a feel as to how the network grows over time. Each r_{out} and r_{in} were computed for cumulative groups of years from 1363 up until the latest year in the data set, with a step size of 13[3] (13 years worth of nodes added to network for each coefficient) and plotted in Figure 5.4a and Figure 5.4b respectively.

Both the assortativity coefficients for each group of years illustrated in Figure 5.4 exhibit interesting behaviour from the late 1860s onwards. Also, the proportion of edges included in the calculation pre-1900s, is too small to deduce statistically significant trends from (see Section 4.2). Due to both these reasons, yearly (step size of one) cumulative r_{out} and r_{in} are plotted in Figure 5.5a and Figure 5.5b respectively, for years inclusive of 1860 up until the latest year in the data set.

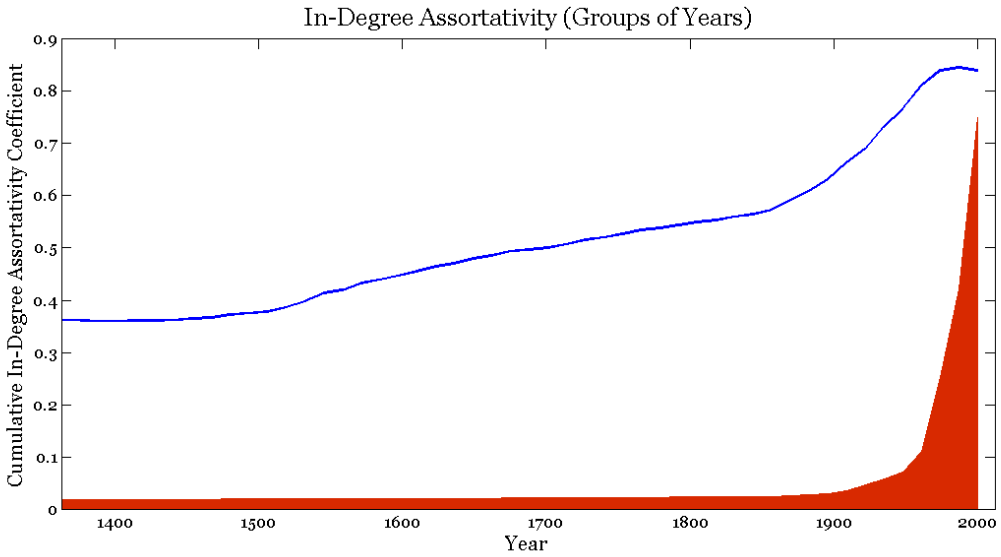
A positive out-degree assotativity coefficient corresponds to assortative mixing by out-degree, and the closer the value of the coefficient is to one, the stronger the tendency for advisers to be connected to an advisee with a similar out-degree. In other words, individuals with many academic siblings go on to have many advisees of their own.

Figure 5.5a indicates that the out-degree assortativity coefficient remains relatively steady at 0.11 from 1860 up until 1900, at which point the coefficient increases to 0.184 in 18 years. From 1918 to 1933, the out-degree assortativity coefficient for the directed network remains steady around a value of 0.185. There is a period of decrease in the out-degree assortativity coefficient between 1934 and 1942, from a value of 0.183 to a value of 0.166, after which it begins to increase steadily to a value of 0.2273 in 1983. The out-degree assortativity coefficeint remains in the range of 0.21 and 0.23 between 1984 up until 2012.

[3]A step size of 13 is chosen so that the period from 1363 onwards is split into 50 year groups, therefore each assortativity coefficient is calcualted 50 times.



(a) Out-degree assortativity coefficient r_{out} (blue) and the proportion of the 138,167 edges included in caluclation (red).

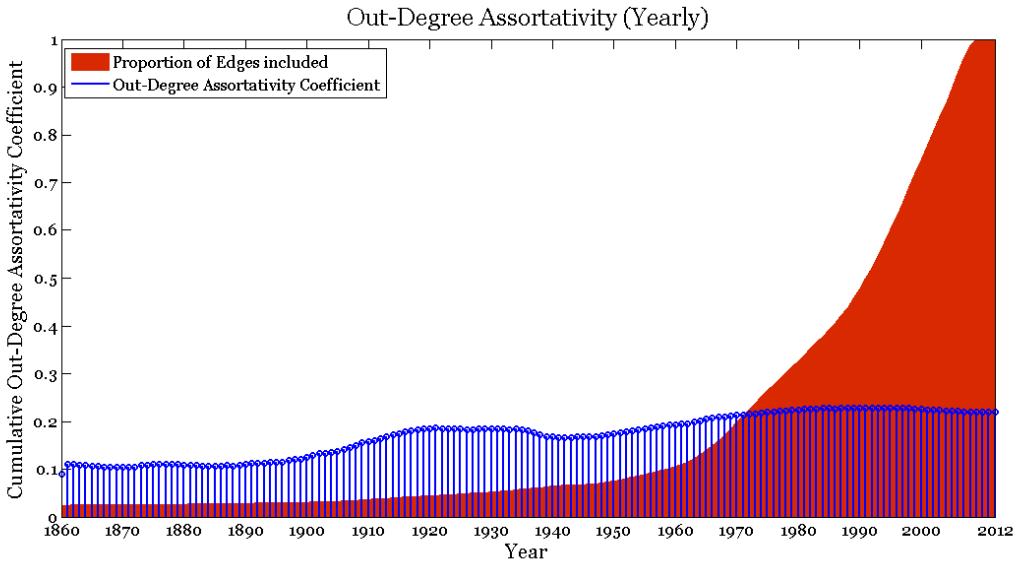


(b) In-degree assortativity coefficient r_{in} (blue) and the proportion of the 138,167 edges included in caluclation (red).

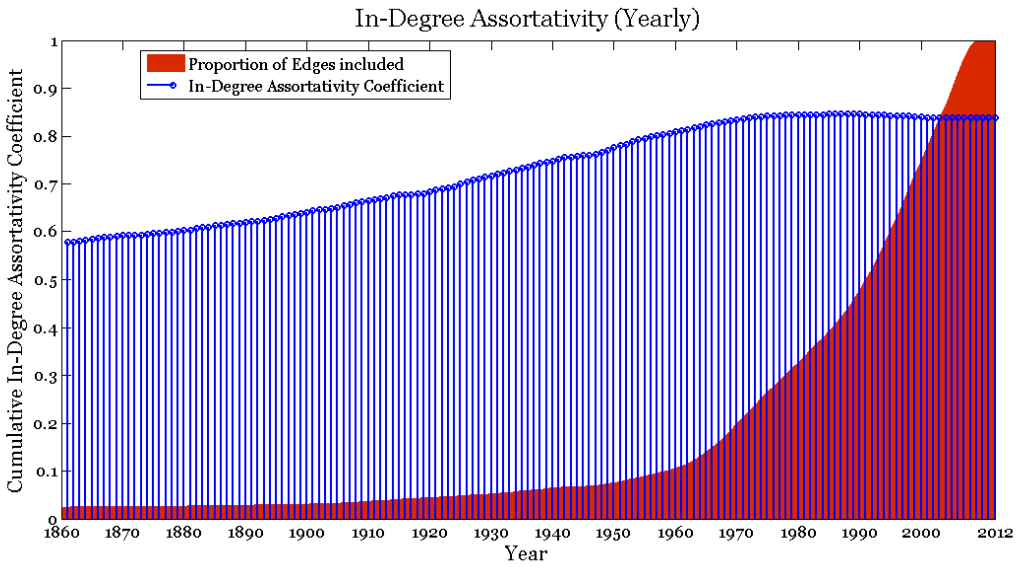
Figure 5.4: Assortativity coefficients as individuals are added to the network per 13 years (1363 - 2012).

A positive in-degree assortativity coefficient close to one implies a strong tendency for advisers to be connected to (i.e. advise) advisees with a similar in-degree. In other words, advisers with a lot of advisers have a tendency to have advisees with a lot of advisers, and advisers with a small number of advisers have a tendency to have advisees also with a small number of advisers.

Although the proportion of edges from the entire directed network seems to grow



(a) Out-degree assortativity coefficient r_{out} (blue) and the proportion of the 138,167 edges included in calculation (red).



(b) In-degree assortativity coefficient r_{in} (blue) and the proportion of the 138,167 edges included in calculation (red).

Figure 5.5: Assortativity coefficients as individuals are added to the network per year (1860 - 2012).

exponentially, from 1860 up until 2009, the in-degree assortativity coefficient plotted in Figure 5.5b increases steadily from 0.57629 in 1860 up until 1970, to a value of 0.83, after which the in-degree assortativity coefficient remains around 0.83.

Chapter 6

Mathematics Genealogy as Undirected Networks

In the directed network representation of the mathematics genealogy tree, which is a directed acyclic graph, the clustering coefficients defined in Section 3.3 cannot be considered, as it contains no cycles, in particular cycles of length of three. However, if we were to make the directed network undirected, then it is possible for loops to form and the clustering coefficients defined in Section 3.3 can be computed. For the purpose of looking at clustering, we consider two types of undirected networks, the undirected counterpart to the directed acyclic graph formed in Chapter 5 and a sibling network, in which individuals are connected to their supervisors as well as their academic siblings by undirected edges. Academic siblings are individuals supervised by the same advisor.

Considering the sibling network also has another advantage. Studying the structural properties of the sibling network and comparing it to that of the undirected network, can also provide us an insight into the interactions between academic siblings, as well as insights into the academic families of 2 generations, consisting of academic parents (the advisers of an individual) and their academic children (advisees of an individual).

6.1 Undirected Genealogy Network

This is exactly the same as the directed acyclic graph representation described in Chapter 5, except that the edges are now undirected instead of directed. The adjacency matrix of the undirected counterpart will be denoted as $\mathbf{U}$ and has elements

$$\mathbf{U}_{ij} = \begin{cases} 1 & \text{if there is an advisor-advisee relationship between } j \text{ and } i, \\ 0 & \text{otherwise.} \end{cases}$$

By construction, $\mathbf{U}$ is symmetric.

This undirected network has a mean degree $c \approx 2.0150$, which is double the mean degree of the directed network given in Chapter 5. Using the degree of the nodes for

which we have information on the year an individual was awarded his/her degree, we can calculate the mean degree of the nodes over time. If we denote Υ_j the set of nodes that were awarded their degree in year j, the mean degree for year j is given by

$$\frac{\sum_{i \in \Upsilon_j} k_i}{|\Upsilon_j|}, \tag{6.1}$$

where we recall that k_i is the degree of node i, and $|S|$ denotes the cardinality[1] of the set S. Figure 6.1 is a plot of the mean degree defined in (6.1) for the undirected network over time. The mean degree over the first few centuries up until the 20th century is very

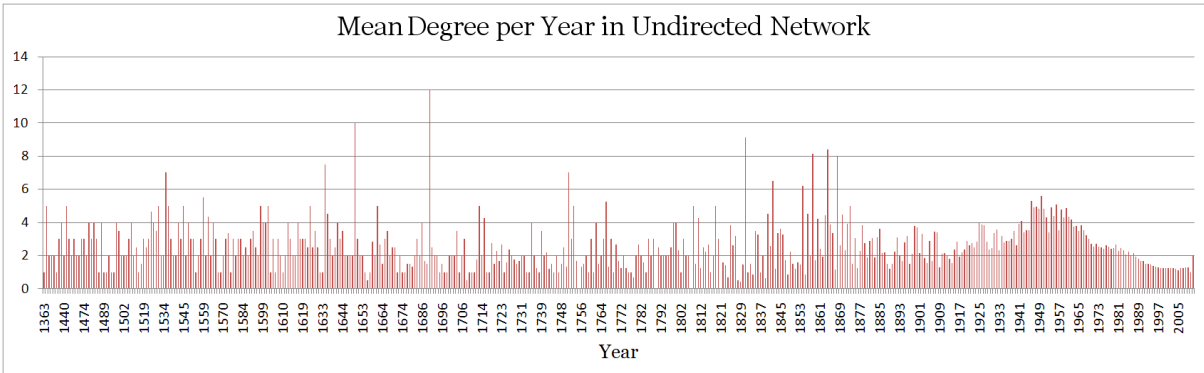


Figure 6.1: Mean degree of nodes in the undirected network over time.

volatile. This can be explained by the few number of nodes over those years for which the mean has been calculated (see Figure 4.1 and Section 4.2). During the 1900s, the average degree in the undirected network increases and peaks at the value between 4 and 6 around the mid 1900s, from after which point it decreases. The average degree continues to decrease through the first decade of the 21st century too, to an average degree of 1 (excluding the last bar in Figure 6.1, which represents the degree of the individual who was awarded their degree in 2012). It is difficult to conclude at this stage if the average degree of an individual is indeed decreasing over the first decade of the 21st century, as the data is still young, in the sense that all individuals that were awarded a degree in the last few decades may not have completed their academic life yet and may not have advised all the students that they might.

6.2 The Sibling Network

An academic sibling of an individual in the mathematics genealogy tree is defined to be another individual who has the same advisor. Here, the sibling network has the same edges as stated for the undirected network given in Section 6.1, but in addition, also has

[1]The cardinality of a finite set is the number of distinct elements in the set.

edges between academic siblings. A small example to illustrate how the edges in the sibling network compares to the undirected (counterpart of the directed acyclic graph) network is given in Figure 6.2, where the extra edges in the sibling network are indicated by in green.

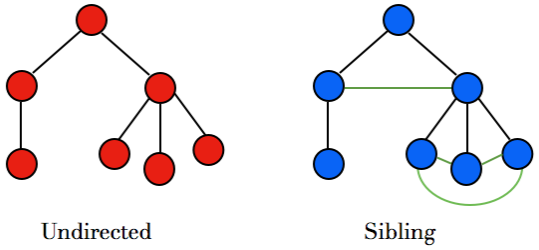


Figure 6.2: Difference between the structure of the two undirected networks considered here, illustrated by a small subset example.

The adjacency matrix of the sibling network will be denoted as $\mathbf{S}$ and has elements

$$\mathbf{S}_{ij} = \begin{cases} 1 & \text{if there is an advisor-advisee or sibling relationship between } j \text{ and } i, \\ 0 & \text{otherwise.} \end{cases}$$

The edges between each sibling can be found by computing all of the combinations of the pairs of nodes that share the same advisor. The adjacency matrix of the sibling network, $\mathbf{S}$, is much denser than the adjacency matrix of the undirected matrix, $\mathbf{U}$ (see Section 6.3). Figure 6.3 is a plot of the mean degree, given by (6.1), over time for the sibling network. The sibling network has a mean degree of 14.8499, significantly larger than the undirected network given in Section 6.1. Just as for the undirected network

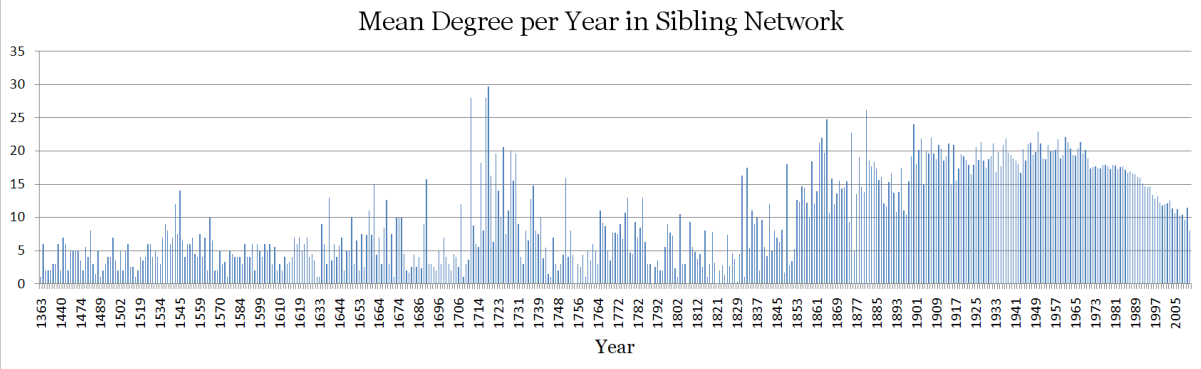


Figure 6.3: Mean degree of nodes in the sibling network over time.

given in Section 6.1, the mean degree of the sibling network given in Figure 6.3 over the first few centuries up until the 20^{th} century is very volatile, due to the few number of nodes over those years for which the mean has been calculated (see Section 4.2). The

mean degree can be seen to oscillate between 15 and 20 over the first two thirds of the 1900s, until 1970. A noticeable drop by 2.5 in the mean degree is observed in 1971, after which the mean degree is fairly constant for a decade with a value of about 17. From 1982 onwards, there is a gradual decrease in the mean degree to a value of approximately 10 in 2010. As for the undirected case in Section 6.1, the decrease observed in the mean degree for the sibling network over the past 4 decades, may not be a trend in time and could be accounted for by the data being young. However, if this is indeed an emerging trend, this suggests that academic families of 2 generations, consisting of parents and children, are getting smaller. The family relations used here, such as parents and children, can be inferred to the academic genealogy, for example academic parents are the advisers of an individual, and academic children are the advisees of an individual.

6.3 Degree Distributions

The undirected and the sibling networks by construction have the same number of nodes, 137,138, as the directed network given in the previous chapter. However, the number of edges, m, in the siblings network differs from the number of edges in the directed network and its undirected counterpart network. The siblings network has 880,109 more edges than the other two networks, which have 138,167 edges. Hence, the sibling network is much denser than the undirected network. Figure 6.4 is a histogram of the degree distribution of the nodes of the undirected network and is plotted on a log-log scale which is included as an inset in the top right corner of the same figure.

Figure 6.4 indicates that the majority of nodes in the undirected network have a degree of 1 or 2, which is in agreement of the degree distribution of the directed network given in Figure 5.2.

Figure 6.5 is a histogram of the degree distribution of the nodes of the sibling network, and is plotted on a log-log scale which is included as an inset in the top right corner of the same figure.

Although the majority of nodes in the sibling network have a degree of 1 which can be seen from Figure 6.5, the degree distribution of the sibling network is more spread out over the other degrees than that of the undirected network, given in Figure 6.4. A significant proportion of nodes have a degree of more than 1. It can be seen that the degree distribution has a significant 'tail' to the distribution, corresponding to nodes with substantially higher degree.

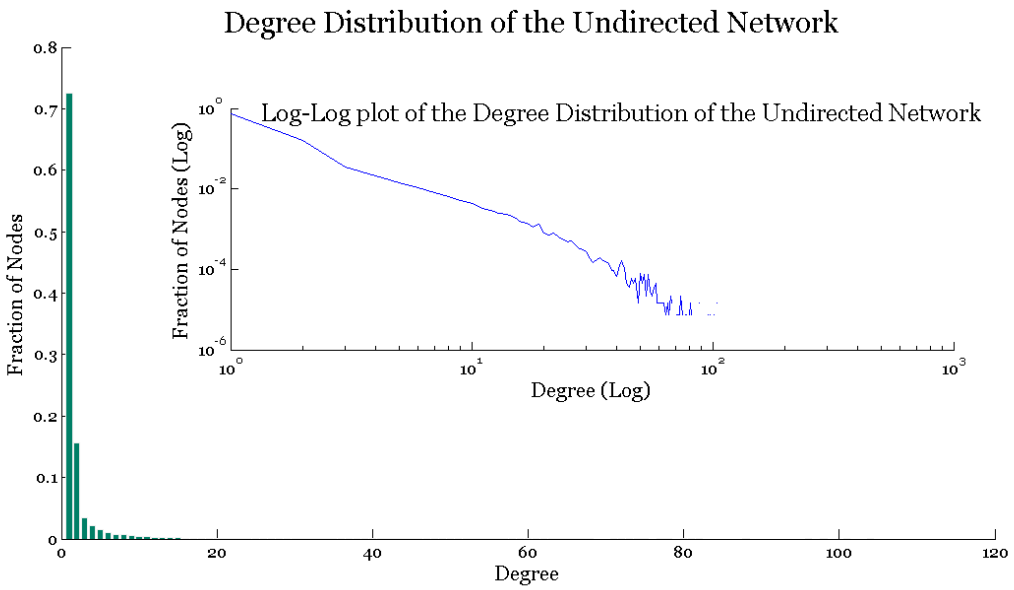


Figure 6.4: Degree distribution of the undirected network.

6.4 Degree Assortativity: Pearson Correlation Coefficient

Although we can use (3.17) to calculate the Pearson correlation coefficient of the degrees, but because the terms in the formula contain summations over edges, we would need to use the edge list. Using the edge list to calculate with requires computing with for loops and this is computationally inefficient in MATLAB. Hence we rewrite each of the terms in (3.17) in terms of the adjacency matrix. Suppose the adjacency matrix of an undirected network is $\mathbf{A}$ and the degree vector, say k (where k_i is the degree for node i), then we can rewrite each term in (3.17) as

$$\begin{aligned}
\sum_{i\in\mathcal{E}} x_i y_i &= \frac{1}{2}\sum_{i,j\in\mathcal{N}} A_{ij}k_i k_j = \frac{1}{2}(\mathbf{A}\ k)^T\ k, \\
\sum_{i\in\mathcal{E}}(x_i + y_i) &= \frac{1}{2}\sum_{i,j\in\mathcal{N}} A_{ij}(k_i + k_j) = \frac{1}{2}\left[2\sum_{j\in\mathcal{N}}\left(\sum_{i\in\mathcal{N}} A_{ij}k_i\right)_j\right] = \sum_j(\mathbf{A}\ k)_j, \\
\sum_{i\in\mathcal{E}}(x_i^2 + y_i^2) &= \frac{1}{2}\sum_{i,j\in\mathcal{N}} A_{ij}(k_i^2 + k_j^2) = \frac{1}{2}\left[2\sum_{j\in\mathcal{N}}\left(\sum_{i\in\mathcal{N}} A_{ij}k_i^2\right)_j\right] = \sum_j(\mathbf{A}\ (k^T\ k))_j,
\end{aligned}$$

where we recall that $\mathcal{E}$ is used to denote the set of edges, and $\mathcal{N}$ to denote the set of nodes. There is a factor of $\frac{1}{2}$ in the above expressions because the adjacency matrix of an undirected network counts every edge twice.

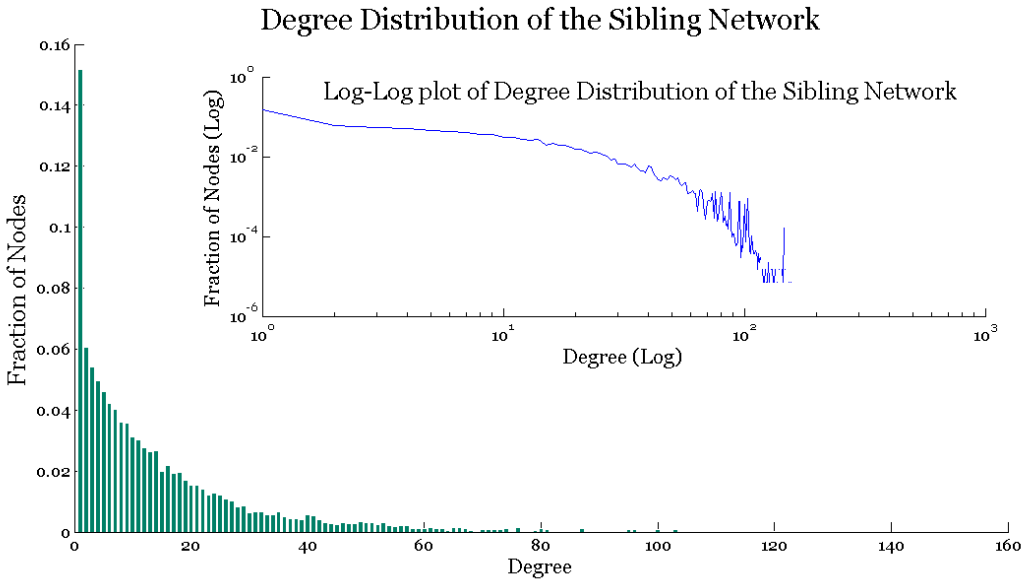


Figure 6.5: Degree distribution of the sibling network.

Hence, the Pearson correlation coefficient given in (3.17) can be written as

$$r = \frac{\frac{1}{2m}(\mathbf{A}\ k)^T\ k - \left[\frac{1}{2m}\sum_j (\mathbf{A}\ k)_j\right]^2}{\frac{1}{2m}\sum_j (\mathbf{A}\ (k^T\ k))_j - \left[\frac{1}{2m}\sum_j (\mathbf{A}\ k)_j\right]^2}, \tag{6.2}$$

where

$$2m = \sum_i k_i. \tag{6.3}$$

The Pearson correlation coefficient calculated using (6.2) for the undirected network is -0.2324 and for the sibling network is 0.8335.

Because a Pearson correlation coefficient $r < 0$ indicates disassortative mixing, this suggests in the undirected network, high-degree nodes have a tendency to attach to low-degree nodes. However, in the sibling network, a positive assortativity coefficient that indicates assortative mixing suggests that high-degree nodes have a tendency to attach to high-degree nodes.

6.5 Clustering Coefficients

The clustering coefficient given by (3.23) requires one to calculate the number of loops of length 3 and the number of paths of length 2 in the network. We can write the number of paths of length 2 in the network in terms of $\mathbf{A}$, the adjacency matrix of the network, as

$$\sum_{\substack{ij \\ i\neq j}} \sum_k A_{ik}A_{kj} = \sum_{\substack{ij \\ i\neq j}} A_{ij}^2 = \|\mathbf{A}^2\| - \mathrm{Tr}(\mathbf{A}^2),$$

because $A_{ik}A_{kj} = 1$ if and only if there is an edge between nodes i and k and between nodes k and j. In this calculation, $i \neq j$ is required, or else this would just be a loop of length 2, i.e one would double count the edges. Similarly, the number of loops of length 3 is

$$\sum_i \sum_{jk} A_{ij}A_{jk}A_{ki} = \sum_i \left[A^3\right]_{ii} = \mathrm{Tr}(\mathbf{A}^3),$$

so we can calculate the clustering coefficient (3.23) from the adjacency matrix of the network by

$$C = \frac{\mathrm{Tr}(\mathbf{A}^3)}{\|\mathbf{A}^2\| - \mathrm{Tr}(\mathbf{A}^2)}. \tag{6.4}$$

In the undirected network, loops of length 3 form only when an academic grandparent and parent together advise a child, as edges only connect advisers and advisees. Paths of length 2 in the undirected network can only occur in the following cases

- when an advisee has strictly more than 1 adviser, or
- when an adviser has strictly more than 1 advisee, or
- when an individual has at least 1 adviser and at least 1 advisee.

The clustering coefficient for the undirected network is $C \approx 0.0060$, which indicates that not many loops of length 3 occur compared to the number of paths of length 2. There are several different cases when paths of length 2 can arise in the undirected network, therefore a clustering coefficient of $C = 0.0060$ for the undirected network, indicates that there are not many loops of length 3 formed. In terms of the mathematics genealogy tree, this network diagnostic suggests that not many advisers and advisees supervise the same individual.

In the sibling network, however, loops of length 3 can arise if an individual has more than one advisor (who are siblings) or more than one advisee, as not only are advisers and advisees connected by an edge, but there are edges between academic siblings. As siblings are also connected, paths of length 2 can only arise in the following cases:

- when an individuals advisers' siblings do not supervise them as well, or
- when an individual has no siblings and at least 2 advisers that are not siblings, or
- when an individual has no siblings, 1 advisee, and 1 one adviser, or
- when an individual has no siblings, 1 advisee and at least 2 advisers that are not siblings.

A clustering coefficient of $C \approx 0.8654$ calculated for the sibling network indicates that there are not many more paths of length of 2 than there are loops of length 3 in the network. The fact that several nodes in the sibling network have a degree in the range of 1 to 20, (from the degree distribution of the network given in Figure 6.5), it is not surprising that there are a large proportion of loops of length 3 present in the sibling network when an individual only needs to have more than 1 advisee or more than 1 adviser who are siblings.

Local Clustering Coefficient over Time

Similarly, we can write the local clustering coefficient (3.24) in terms of $\mathbf{A}$, the adjacency matrix of the network, and is given by

$$C_i = \frac{A_{ii}^3}{\left(\sum_j A_{ij}^2\right) - A_{ii}^2}. \tag{6.5}$$

We define the local clustering coefficient in order to study clustering in both the undirected networks over time. We can do this because the local clustering coefficient is defined for each individual in the data set, so we can associate a year with each coefficient that we calculate, whereas this is not possible for the clustering coefficient given in (6.4). For each individual, we calculate the local clustering coefficient using (6.5) and then group the coefficients by the year the individual was awarded their degree. The mean local clustering coefficient is plotted for different year groupings for both the networks in Figure 6.6. If we denote $\mathfrak{C}_j$ for the set of nodes that were awarded their degree in the j^{th} century, the mean clustering coefficient for century j is given by

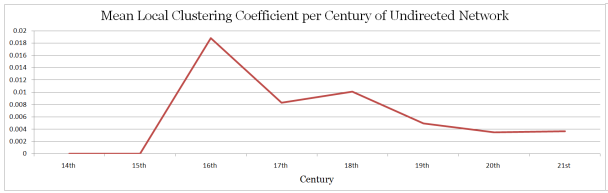
$$\frac{\sum_{i \in \mathfrak{C}_j} C_i}{|\mathfrak{C}_j|}, \tag{6.6}$$

where C_i denotes the local clustering coefficient for node i. The mean clustering coefficient for each century the data is available for, is plotted in Figures 6.6a and 6.6b for the undirected and sibling networks respectively. If the set of nodes that were awarded their degree in year j is denoted as Υ_j, the mean clustering coefficient for year j is given by

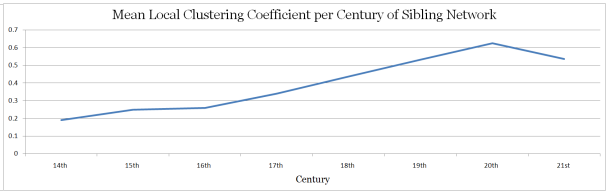
$$\frac{\sum_{i \in \Upsilon_j} C_i}{|\Upsilon_j|}. \tag{6.7}$$

Figures 6.6c and 6.6d are plots of the mean clustering coefficient (6.7) for each year starting from the earliest year in the data set (1363), for the undirected (red) and sibling (blue) networks, respectively. A large variance in the mean local clustering coefficients pre-1900s is due to the small number of individuals it has been averaged for per year, and

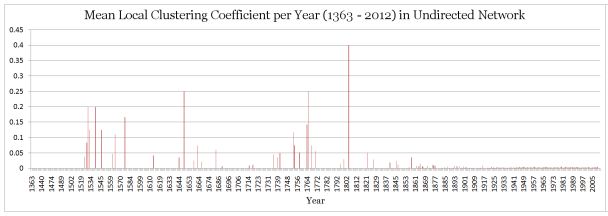
so statistically significant trends cannot be deduced from these numbers (see Section 4.2). Figures 6.6e and 6.6f are plots of the mean clustering coefficient for each year starting from 1860, for the undirected and sibling networks respectively.



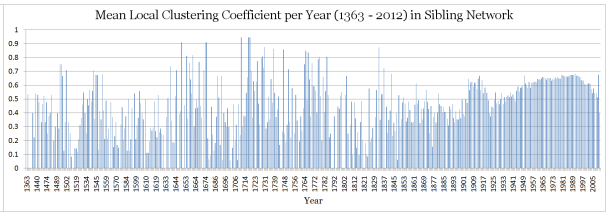
(a) Undirected network: Mean C_i per century



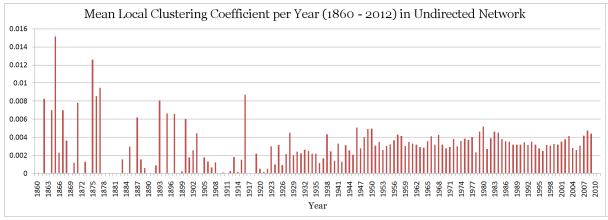
(b) Sibling network: Mean C_i per century



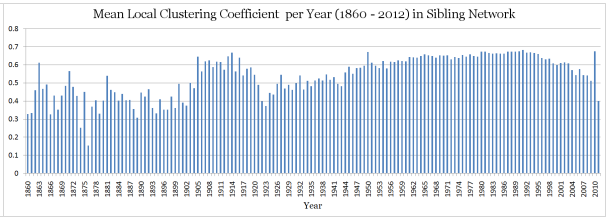
(c) Undirected network: Mean C_i per year



(d) Sibling network: Mean C_i per year



(e) Undirected network: Mean C_i per year from 1860



(f) Sibling network: Mean C_i per year from 1860

Figure 6.6: Mean local clustering coefficient, C_i over time for the undirected and sibling network.

From Figure 6.6a, one can see that the mean local clustering coefficient for each century peaks in the 16^{th} century for the undirected network with a value of 0.0188, after which it continues to decrease per century until the 20^{th} century. The mean local clustering coefficient for the 21^{st} century is relatively the same as for the previous century in the undirected network. However, for the sibling network, the mean of the local clustering coefficient taken from each century increases every century from the 14^{th} century to the 20^{th} with a value of 0.625, after which it decreases for the current century to 0.536. Any 'trend' stated for the period prior to the 20^{th} century should be treated with caution as these figures are not statistically significant, due to the small number of individuals these coefficients have been calculated for.

Examining the local clustering coefficients averaged per year for the undirected network, in Figure 6.6c, and for the sibling network given by Figure 6.6d, we can see that these coefficients are very volatile from 1363 up until the early 1900s. This volatility can

be explained by the small number of individuals in each year group that the local clustering coefficient has been averaged for, (see Figure 4.1 and Section 4.2). For this reason, we restrict our attention to the 20^{th} and 21^{st} centuries in Figures 6.6e and 6.6f. From Figure 6.6e, we can see that the mean local clustering coefficient per year for the undirected network oscillates roughly every 5 years between values of 0.003 to 0.006, which is very small in magnitude compared to the mean clustering coefficients for the sibling network. The mean clustering coefficient for the sibling network for each year, indicated in Figure 6.6f is fairly stable around a value of 0.6 during the 20^{th} and 21^{st} centuries, with a noticeable dip that occurs between the years 1920 and 1960, where the decrease happens in the first 3 years to a value of 0.4, and gradually increases back up to 0.67. Towards the late 1990s, and into the beginning of the 21^{st} century, the mean local clustering coefficient gradually begins to decrease to a value of 0.5 observed in 2009, the last year that has a number of nodes greater than a 1,000. It might not be reliable to deduce trends from the mean local clustering coefficients calculated for nodes who received their degree after 2009 due to the low number of nodes the information is available.

Chapter 7

Conclusions

The Mathematics Genealogy Project data was represented as a mathematics genealogy tree, to explore the advisor-advisee relationship in the community of mathematicians and infer from this the influence an advisor has on their advisees' supervising behaviour, or how the advisees are influenced by their advisers.

We analysed the MGP as a data set to try and understand its main characteristics in Chapter 4. The mathematics genealogy tree was modelled by 3 different networks, and various network diagnostics were calculated in Chapters 5 and 6 to better understand the advisor-advisee relationship in the community of mathematicians, and possibly infer from this how advisees are influenced by their advisers.

Observations in the MGP data set

Exploring the MGP data set itself, in Chapter 4, without using network theory, gave us an insight into how the mathematical community has developed, and how new members join the family of mathematicians.

An individual in the data set was found to have up to a maximum of 5 advisers and a maximum of 103 advisees. This indicates that in the mathematics community, it is possible for individuals to have many more advisees than advisers.

The number of degrees awarded and the number of advisers an individual had over time[1] was analysed.

The number of mathematical degrees that is awarded over time can be used to quantify and measure the growth of the community of mathematicians. Analysis showed that the majority of individuals in the MGP data set were awarded their degree in the 20^{th} century and the first tenth of the 21^{st} century. The number of degrees awarded per year started increasing in the late 1800s with a massive increase in growth rate beginning in the mid

[1]Some individuals in the data set do not have a year associated with them, so this result is based on the individuals who had information listed on the year they were awarded their degree.

1900s, that still continues up until today. The years from 1860 to 1959 was a period of steady growth of the mathematics genealogy tree, except for a dip in the number of degrees awarded we observed over the post-World War II years of 1944 to 1947, suggesting a pause in growth of the community during this 3 year period. In the 1970s, the number of individuals awarded a degree per year remained relatively constant, until 1985, when the numbers began to climb rapidly and continued to increase throughout the 1990's. The year in which the most number of individuals in the data set were awarded a degree was observed to be in 2005. However shortly after 2005, there was a significant decrease. One could speculate that this could be due to the recent downturn in the economy. The number of individuals awarded a degree in the 2010s decreased sharply from 2009, indicating that the data for the 2010s is incomplete.

In Chapter 4 we also observed that the maximum number of advisers an individual had, increased in the past 3 centuries. However, in each of the 19^{th}, 20^{th}, and 21^{st} centuries, the majority of individuals ($\sim 80\%$) had just 1 advisor.

Observations using Network Theory

By modelling the Mathematics Genealogy Project data set of 137,138 individuals as a network and computing network diagnostics, we are able to better understand the advisor-advisee relationship, and possibly infer from this how the advisees are influenced by their advisers.

The Directed Network

The first network we constructed was the directed network in Chapter 5, in which the nodes were taken as the individuals and the directed edges to be the direction of advice in the advisor-advisee relationship. The plot of the in- and out-degree distributions and the mean degree of the directed network in Chapter 5 just reconfirmed that the majority of individuals in the MGP data set have 1 advisor, and also that the maximum number of advisers an individual had was 5 and the maximum number of advisees an individual had was 103. However these network diagnostics also indicated that both of the extreme cases of 5 advisers and 103 advisees are rare. The out-degree distribution plot also indicated that the majority of advisers have just 1 advisee.

We then calculated the degree assortativity diagnostic on the directed network, in the hope to better understand the advisor-advisee relationship, and infer from this the influence the advisor has on their advisee, in terms of their advising behaviour. Calculations indicated a slight assortative mixing by out-degrees, which, in terms of the mathematics genealogy tree and the advisor-advisee relations, implies that advisees who have many

academic siblings have a tendency to go on to become advisers with many of their own students. Also, advisers who advise a small number of students have a slight tendency to influence their advisees to have small number of students themselves. The in-assortativity coefficient calculated for the directed network, indicated a strong assortative mixing by in-degree. In terms of the adviser-advisee relations, this suggests that advisers who were supervised by many individuals have a tendency to advise a student with other individuals, or that advisers who were supervised by few individuals have a tendency to advise students by themselves or only with few other individuals. But closer inspection of the in-degree distribution matrix indicated that this result was mainly driven by the second case. In which case, individuals who were supervised by just 1 adviser had a tendency to advise their students by themselves.

These tendencies seen for the directed network as a whole, was observed to increase in strength, as more individuals were added to the mathematics genealogy tree over time.

The Undirected Networks

Unable to consider clustering (as defined in Section 3.3) for the directed network, the mathematics genealogy tree was then modelled by two undirected networks.

Interested to see the interactions between academic siblings, we constructed the sibling network, in Chapter 5. The nodes of the sibling network were taken as the individuals and the edges to be between advisers and advisees and also between academic siblings. However diagnostics calculated from this could not be directly compared to the directed network, therefore we also constructed the undirected counterpart to the directed network in tandem.

Studying the structural properties of the sibling network and comparing it to that of the undirected network, can provide us an insight into the interactions between academic siblings, as well as insights into academic families of 2 generations, consisting of academic parents (the advisers of an individual) and their academic children (advisees of an individual).

The mean of the total number of advisers and advisees an individual had, was found to be 2, whereas the mean of the total number of advisers, advisees and academic siblings and individual had, was found to be 15. These results suggest that the mean number of siblings an individual in the MGP data set has is 13. The mean of the total number of advisers and advisees an individual had was analysed per year (mean degree per year in the undirected network). This analysis showed that the mean number of advisers and advisees in total an individual had, increased during the 1900s and peaked at the value between 4 and 6 around the mid 1900s, from which point it decreased through the first decade of the 21st century, to an average of 1. However it is difficult to conclude at this

point in time if the average number of advisers and advisees an individual had, is indeed decreasing over the first decade of the 21st century, as the data is still young. It is young in the sense that all individuals that were awarded a degree in the last few decades may not have completed their academic life yet, and may not have advised all the students that they might.

The mean of the total number of advisers, advisees, and academic siblings an individual had was also analysed per year (mean degree per year in the sibling network). This analysis showed that there was a noticeable drop in the mean of the total number of advisers, advisees, and academic siblings an individual had, by 2.5, to a value of 17 observed in 1971, after which the mean remains fairly constant up until 1982. A decrease seen in the mean of the total number of advisers, advisees and academic siblings an individual had over the past 4 decades, might not be a trend in time and could be accounted for by the data being young. However, if this is indeed an emerging trend, this suggests that academic families of 2 generations, consisting of parents and children, are getting smaller. Individuals have fewer advisers, advisees and academic siblings in total than previously seen.

The degree distribution of the undirected network indicates that the majority of the individuals in the mathematics genealogy tree have a total of 1 or 2 number of advisers and advisees combined. The degree distribution of the sibling network shows that the majority of individuals have up to a total of 20 number of advisers, advisees and academic siblings combined.

The Pearson correlation coefficient of the degrees calculated for the undirected network indicates that the undirected version of the directed network is marginally disassortative. We can interpret this result in terms of the mathematics genealogy tree and advisor-advisee relations as individuals who have many advisers and advisees, have a tendency to be connected to individuals with less number of advisers and advisees, where the connection here can be taken as both to receive and give advice.

The Pearson correlation coefficient of degrees calculated for the sibling network indicates a strong assortative mixing by degree, quite the opposite result for the undirected network. This result indicates that individuals in the data set have a tendency to supervise or to be supervised by other individuals that have a similar total number of advisers, advisees and academic siblings. With this diagnostic by itself, it is difficult to understand if the sibling network is indeed suggesting an interesting structure of the mathematics genealogy tree, or that adding the extra edges between academic siblings has lost too much information to be able to extract something useful using this diagnostic.

The clustering coefficient calculated for the undirected network, indicated that there is not much transitivity in the network, i.e. not many loops of length 3 formed, which

occurs only when advisees pair with their advisers to supervise another individual in this network. In terms of the mathematics genealogy tree, this network diagnostic suggests that not many individuals advise students with their own advisers. The frequency of this type of advising behaviour was found to be fairly constant over time.

The clustering coefficient calculated for the sibling network, however, indicates almost perfect transitivity in the network, i.e. many loops of length 3 form in comparison to the number of paths of length 2. In terms of the mathematics genealogy tree, a loop of length 3 in a sibling network can arise when an individual in the data set has more than 1 advisee or more than 1 adviser who are siblings. The fact that several individuals in the data set have up to a total of 20 number of advisers, advisees and academic siblings, it is not surprising that an individual would have more than 1 advisee or 1 adviser who are siblings (at least 2 academic parents of an individual that were advised by the same academic grandparent of the individual). This leaves us to question, once again, if adding extra edges between siblings to the undirected network loses too much of the structure of the mathematics genealogy tree for the purpose of this diagnostic.

Chapter 8

Discussions

By modelling the Mathematics Genealogy Project data set of 137,138 individuals as a network and computing network diagnostics, we are able to better understand the advisor-advisee relationship, and possibly infer from this how the advisees are influenced by their advisers. Some of the main results mentioned in Chapter 7 are discussed and possible explanations are hypothesised here.

In-degree Assortativity

Using various degree assortativity diagnostics gives us an insight into the difference in the number of advisers and/or the number of advisees between two nodes connected in a network, as discussed in Chapter 7. It was shown in Chapter 5 that there is a very strong assortative mixing by in-degrees in the directed network, where the in-degree of a node in the directed network represents the number of advisers an individual has in the mathematics genealogy tree. In other words, this result indicates that individuals have a strong tendency to connect to other individuals with the same number of advisers, and from computing the in-degree assortativity coefficient per year, as more individuals were added to the network, this tendency has been observed to generally increase over time. There are possible explanations for advisees to have a tendency to go on and advise similar numbers as their own advisers. One could hypothesise that the advisees agree with their advisers on the number of supervisors a student needs which might be necessary in their line of research area. This is assuming that individuals are advised by advisers interested in a similar research area as the subject area of their degree. For example, in the case of high in-degrees, i.e. when individuals are supervised by many advisers, the research area of an individual could require them to study a number of subject areas and may imply that individuals need advice from experts in different fields. Alternatively, in the case of low in-degree, i.e. the case when individuals have few advisers, the subject areas might be very specialised and focused, and might only need or get advice from very

few advisers. If one had the Mathematics Subject Classification numbers to classify the subject area associated with each node, one could explore this possibility.

Out-degree Assortativity Versus Degree Assortativity of the Undirected Network

The calculations in the Chapter 5 suggest that there is a slight assortative mixing by out-degrees in the Directed network.

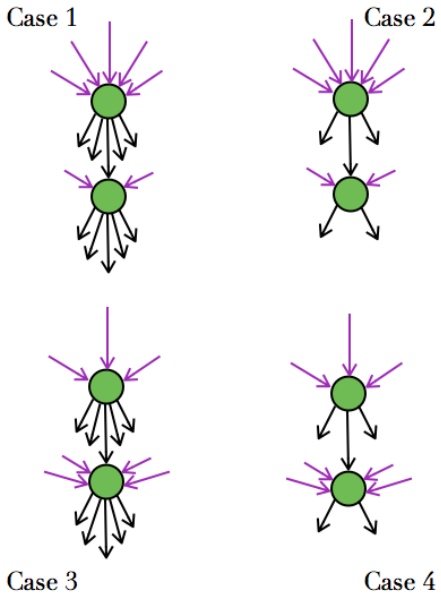


Figure 8.1: Black edges represent the out-degree, and the purple directed edges represent the in-degree added.

In Chapter 6, however, the Pearson correlation coefficient of the degrees calculated for the undirected network indicates that the undirected version of the directed network is marginally disassortative, i.e. high-degree nodes have a slight tendency to attach to low-degree nodes. We consider the cases when these results coincide. We note that by adding the in-degrees for each node, in effect, we can make the directed network, which is slightly assortative by out-degrees, into an undirected network which is slightly disassotative by degrees. The different cases of how this might happen is indicated in Figure 8.1. The black edges are the edges in the directed network that point to a node (edges that point away from the nodes are not considered in the calculation of r_{out}). The purple edges can be interpreted to represent the edges that need to be added to make the undirected network slightly disassortative by degree. Because all of the situations given in Figure 8.1, indicated by the calculations of the out-degree assortativity coefficient of the directed network and the Pearson correlation coefficient of the undirected network, cover several of the possible combinations of in- and out-degrees of an advisor and an advisee in the directed network, not much can be deduced from this on its own. Further analysis needs to be done to understand which of these cases indicated in Figure 8.1 is most common in the mathematics genealogy network, and also what the driving factor is. For example one could hypothesise that in some subject areas, an individual would only need 1 or 2 advisers, and explore how the number of advisers for each subject area changes over time.

Another interesting point to note is that the near to perfect transitivity found in the sibling network could be driven by individuals with many advisees as a result of cases 1 and 2, given in Figure 8.1, found at large in the mathematics genealogy directed network.

Chapter 9

Further Work

9.1 Assortativity Using Other Characteristics

We have only considered the special case of assortative mixing by degree. Although the data of other characteristics for each individual in the Mathematics Genealogy Project data set is limited (Table 4.1), one can still construct a smaller network using the subset of nodes for which that data is available.

Assortative mixing by the location and university name at which the degree of an individual was awarded, is considered here separately, for all three networks, the directed (DAG), undirected and sibling networks and the results are can be found in Table 9.1.

Node Label	DAG	Undirected	Sibling
Country	0.7007	0.6995	0.8955
University	0.1251	0.1235	0.7292

Table 9.1: Assortativity Coefficient

The numbers given in Table 9.1 are calculated using (3.10). If either node at the ends of an edge has no label information (blank fields in the data), then the edge is excluded from the mixing matrix and hence the computation.

The numbers in Table 9.1 indicates a high assortative mixing by the country from which the degree is awarded, for all three networks. The assortativity coefficients for mixing by university for the directed and undirected networks indicates only a slight assortative mixing (assortativity coefficients are close to zero), whereas mixing by university that awarded the degree is highly assortative in the Sibling network. This could indicate that although advisers and advisees might not have a strong tendency to have been awarded degrees from the same university, but academic siblings have a strong tendency to be awarded a degree from the same university. This suggests that advisers might

have a strong tendency to advise their advisees in the same universities, which is perhaps worth exploring in further detail.

Interesting as these results may be, however, this approach does not give a full picture of the whole family of mathematicians, as missing information would imply excluding edges in diagnostics. Consider the small example network in Figure 9.1, in which the nodes coloured green have missing type labels. If any of the nodes in an edge do not have a type label associated to them, they are not included in the mixing matrix, and hence diagnostic calculation. Therefore the edges that will not be included as a result of missing node label, are indicated by a dashed blue line in the figure. As a result, the assortativity coefficient might not give an accurate picture of the actual structure of the mathematics genealogy. Ideally, if one can obtain the missing fields[1], so that the data set is nearly

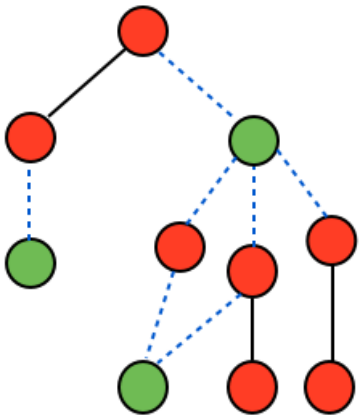


Figure 9.1: A small example network with missing node type labels.

complete, then one can delve further into the tendency patterns and associative behaviour of individuals in the mathematical community by using assortativity mixing for initial probing of the data. Mixing with subject area can be used to trace and understand the evolution of research groups, but it can also be used to help explain the trends discussed in the conclusion. For example, in the conclusion, it was mentioned that a possible reason for an assortative mixing by in-degree in the directed network was due to the subject area an individual studied. One can try and develop new diagnostics for each of the three networks, to explore if there is a correlation between the subject area an individual studies and the number of advisees they have (the out-degree of the individual). One could also explore how this correlation changes over time.

[1]In Appendix A, we discuss our attempt to obtain the full list MSC numbers for subject area is discussed.

9.2 Community Structure

A network is said to show *community structure* if there are groups of nodes that are densely connected to each other, referred to as *communities*, but sparsely connected to other densely connected groups [5]. According to [6], *community detection* is the search for naturally occurring communities regardless the size of the network. In the past 9 years there has been an extensive amount of research in community detection in networks [8], and several methods to detect communities have been proposed. However, one would need to generalise the notion of community structure to work on a directed acyclic graph, as this has not yet been done. It would be particularly interesting to identify any such communities in the mathematics genealogy network, as we can start explore characteristics of communities. For example we could investigate if communities form mainly for people with similar research areas, or location from which the degree is obtained. This area opens up numerous in-leading opportunities to investigate the structure of the mathematics genealogy network and the traits of the communities of mathematicians.

Appendix A

Change in Dissertation Topic

A.1 Original Proposal

The original title of this Dissertation was "Mathematical Genealogy and the Evolution of Mathematical Research Groups", and the aim was to understand how different subject areas of mathematical research have evolved in time. The proposed method was similar to the approach taken for this dissertation (described in Chapter 1): first represent the Mathematics Genealogy Project data as a genealogy network, classify the individuals in research groups, and then use network analysis and various computations to understand the evolution of research groups.

Mathematics Subject Classification Number as Subject Area

The Mathematics Subject Classification (MSC) number[1] is an alphanumeric classification scheme used to classify items in the mathematical literature to help users find the items of present or potential interest to them. This list has been produced and updated jointly by the editorial staffs of Mathematical Reviews and Zentralblatt fur Mathematik in collaboration with the mathematical community. This classification scheme is commonly used by many mathematical journals, in products derived from the Mathematical Reviews Database (MRDB) such as MathSciNet and in Zentralblatt MATH (ZMATH). The Mathematics Subject Classification is a hierarchical scheme[2] with three levels of structure: the first (top) level is represented by a two-digit number, the second by a letter, and the third (most specific) by another two-digit number. Currently, there are 97 first-level classification groups in total.

The Mathematics Genealogy Project data set has the top level, i.e. the first two digits of the Mathematics Subject Classification number for the thesis, which could be used to classify the subject that the individual studies. (See the bar chart in Figure A.1

[1] `http://www.ams.org/mathscinet/msc/msc2010.html`
[2] `http://en.wikipedia.org/wiki/Mathematics_Subject_Classification`

for the distribution of Mathematics Subject Classification numbers in the Mathematics Genealogy Project data set). However, this data available in the Mathematics Genealogy Project data set is limited, as 66% of all the individuals listed in the Mathematics Genealogy Project database do not include a Mathematics Subject Classification number of their thesis.

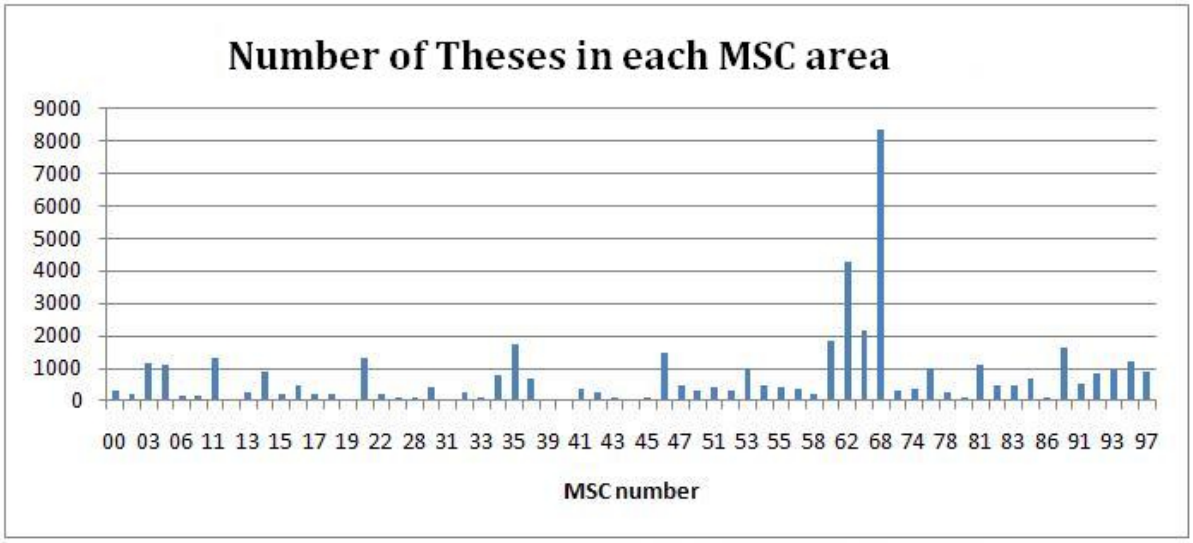


Figure A.1: Mathematics Subject Classification (MSC) number available in the Mathematics Genealogy Project (MGP) data set.

MathSciNet: Source for MSC Numbers

The Mathematical Genealogy Project data set contains the MR Author ID for each individual. The MR Author ID[3] is a unique identifier assigned to each author in the MRDB. The MR Author ID can be used to find each individual from the Mathematical Genealogy Project in the MathSciNet database[4]. This is useful, as the MathSciNet lists the subject areas of the publications of each MR Author ID in the 'Author Profile' pages. Figure A.2 is a screen shot of an example of such an 'Author Profile' page. One can see from Figure A.2 that the subject areas are listed under the heading 'Publication (by number in area)' and have different font sizes. From the html code on which the 'Author Profile' pages in the MathSciNet site are based (see the example in Figure A.3), one can deduce that the font size of each area is dependent on the number of publications of the individual in that area. Also the subject areas are categorised using the MSC numbers. Based on examining a few examples of mathematicians I am familiar with, I have concluded that the number of publications seems to be understated, but that, the

[3] `http://www.ams.org/publications/math-reviews/mr-authors`

[4] MathSciNet is an electronic searchable database, available online, that has both the reviews and citation information of mathematical research publications.

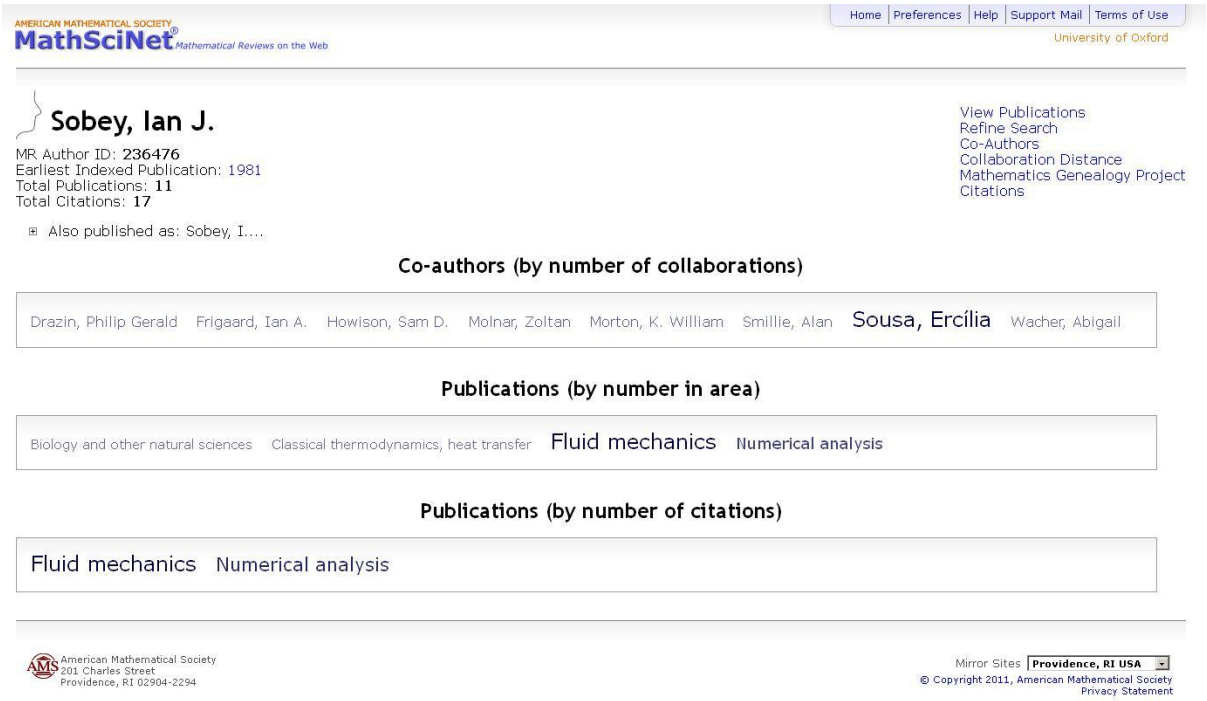


Figure A.2: Screen shot of an 'Author Profile' on MathSciNet.

dominant area of research still seems to be correct. Hence, it is reasonable to take the research group assigned to each individual in the Mathematics Genealogy Project data set, as the area in which they have produced the most publications, i.e. the area with the largest font size.

A.2 Progress Made

In order to obtain the Research Groups classification, the list of MR Author IDs for each individual in the Mathematics Genealogy Project was obtained, and an automated data scraper using Python, using the BeautifulSoup package was written by me. The scraper uses the MR Author ID to find and retrieve the html code for their corresponding 'Author Profile' web page on MathSciNet, as the url[5] is the same for each individual, only differing in the MR Author ID. Using the BeautifulSoup package, the scraper then parses the html code, making it easier to obtain the relevant information which is stated under the heading 'Publication (by number in area)'. It then extracts the two-digit MSC number, the font size, and the publication count (see Figure A.3).

[5] `http://www.ams.org/mathscinet/search/author.html?mrauthid=`*MR Auth ID*

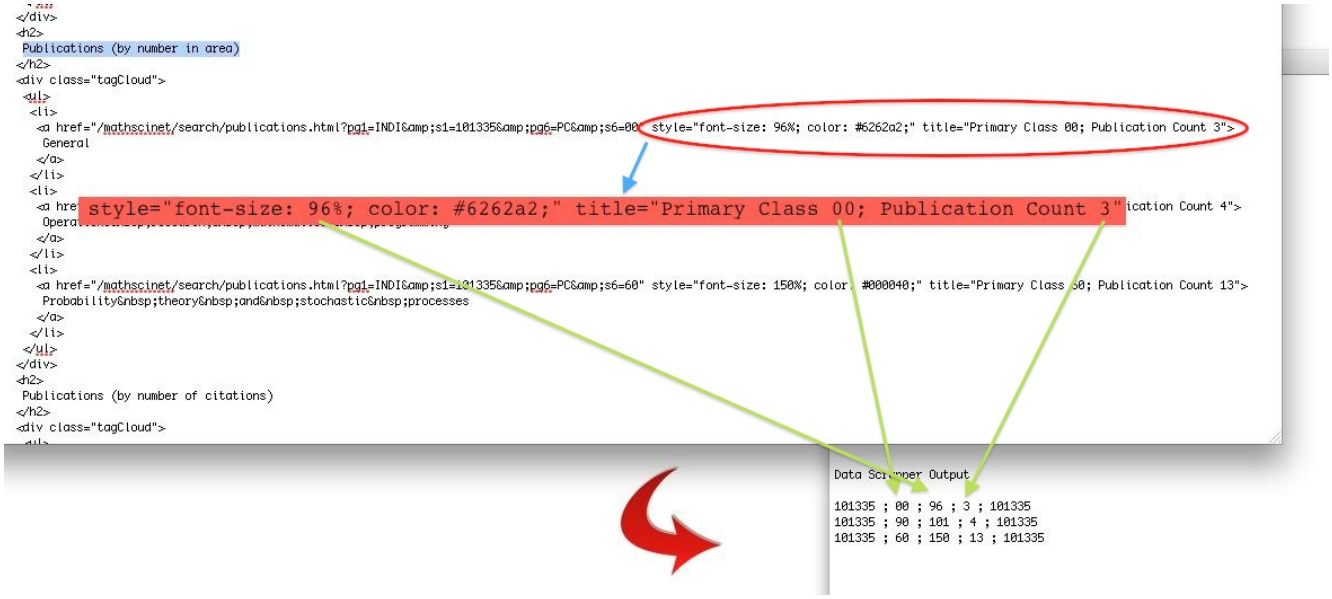


Figure A.3: The data scraper: The top window is the html code retrieved by the data scraper, and the bottom window is the output file. In the html code, the data scraper finds the code circled in red, (highlighted in red is the same text but zoomed in). The scraper then saves the the MR Author ID from input list, the MSC number, font size, publication count, and the MR Author ID in the html code into in an output file (indicated by green arrows).

A.3 Reason for Change

In total, there were approximately 82,000 individuals for whom the data scraper was obtaining information. After acquiring data for almost half of the individuals in the list, my access to the MathSciNet was removed. After which I was requested by the OUCS[6] and the OxCERT[7] to stop scraping data by and delete all acquired data on behalf of the AMS[8], hence the reason to change the focus of dissertation.

[6]Oxford University Computing Services.

[7]Responsibility for network security within the University of Oxford lies with the Oxford University Computer Emergency Response Team (OxCERT).

[8]American Mathematical Society, of which MathSciNet is a service they provide.

Appendix B

Expectation and Standard Deviation of Discrete Distributions

Consider a discrete distribution $\{X_k\}$. The mean μ is equal to the expectation of the distribution, $E[X]$, and is given by

$$\mu = E[X] = \sum_k k\ X_k. \tag{B.1}$$

The variance of the distribution is calculated using the following expression:

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2, \tag{B.2}$$

where

$$E[X^2] = \sum_k k^2\ X_k. \tag{B.3}$$

The standard deviation is then

$$\sigma = \sqrt{\mathrm{Var}[X]}. \tag{B.4}$$

Appendix C

Summary of Results

Network Properties and Diagnostics		Directed Acyclic Graph	Undirected Counterpart	Sibling
Total nodes	n	137,138	137,138	137,138
Total edges	m	138,167	138,167	1,018,246
Mean degree	c	1.0075	2.0150	14.8499
Pearson correlation coefficient	r	$r_{in} \approx 0.8373$ $r_{out} \approx 0.2188$	-0.2324	0.8335
Assortativity coefficient for country	r_c	0.7007	0.6995	0.8955
Assortativity coefficient for school	r_s	0.1251	0.1235	0.7292
Clustering coefficient	C	/	0.0060	0.8654

Table C.1: Basic network statistics for each entire network (1363 - 2012).

Bibliography

[1] R. Malmgren, J. Ottino, and L. A. N. Amaral. The role of mentorship in proteégé performance. *Nature*, 465(7298):622–626, 2010.

[2] S. Myers, P. J. Mucha, and M. A. Porter. Mathematical genealogy & department prestige. to appear in Chaos (Gallery of Nonlinear Images), 2011.

[3] M. E. J. Newman. Assortative mixing in networks. *American Physical Society*, 89(20):5, 2002.

[4] M. E. J. Newman. Mixing patterns in networks. *American Physical Society*, 67(2):026126, 2003.

[5] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[6] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[7] M. Piraveenan, M. Prokopenko, and A. Zomaya. Assortative mixing in directed biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 99, 2010.

[8] M. A. Porter, J-P. Onnela, and P. J. Mucha. Communities in networks. *American Mathematical Society*, 56(9).