Network Theory and CAD Collections

Esmé Frances Louise Anderson

A thesis submitted for the degree of Doctor of Philosophy
Institute for Materials and Processes
University of Edinburgh
May 2016

Abstract

Graph and network theory have become commonplace in modern life. So widespread in fact that most people not only understand the basics of what a network is, but are adept at using them and do so daily. This has not long been the case however and the relatively quick growth and uptake of network technology has sparked the interest of many scientists and researchers. The Science of Networks has sprung up, showing how networks are useful in connecting molecules and particles, computers and web pages, as well as people. Despite being shown to be effective in many areas, network theory has yet to be applied to mechanical engineering design.

This work makes use of network science advances and explores how they can impact Computer Aided Design (CAD) data. CAD data is considered the most valuable design data within mechanical engineering and two places large collections are found are educational institutes and industry. This work begins by exploring 5 novel networks of different sized CAD collections, where metrics and network developments are assessed. From there collections from educational and industrial settings are explored in depth, with novel methods and visualisations being presented.

The results of this investigation show that network science provides interesting analysis of CAD collections and two key discoveries are presented: network metrics and visualisations are shown to be effective at highlighting plagiarism in collections of students' CAD submissions. Also when used to assess collections of real world company data, network theory is shown to provide unique metrics for analysis and characterising collections of CAD and associated data.

Lay Summary

Networks are an ordinary part of modern life; we have never been better connected. The most commonly known is the internet, one giant network; people are linked through social networks, allowing friends and family to be close despite distance, and we are taught to 'network' at work via canapés and polite chat. Networks have been used in science too, linking together particles and molecules, showing that connections can be seen nearly everywhere. Despite this, a large area that is unconnected is mechanical engineering design, specifically Computer Aided Design (CAD).

In this research the links between 3D CAD models in mechanical engineering are explored. It's possible to link CAD models using information about how they are assembled or by how similar they are geometrically; a model of a football and a model of a Buzz Lightyear toy don't have much in common, but a model football and a model tennis ball are more similar. If that similarity is used to build a network, where the toys are the points (nodes) and the link of similarity is a line (edge) between them, a group of models can be turned into a network. And this makes it possible to explore the ways network theory can better connect mechanical engineering design.

This research has taken collections of CAD models from education, students' designs of callipers and steering wheels, and CAD models from industry, engineers' designs of screws, nuts, bolts, and much more complex things, and turned them into networks to be explored. The results of this show that networks can be used to detect plagiarism in students' work at university and that networks can show new information about company CAD collections.

Declaration of Originality

I hereby declare that the research recorded in this thesis and the thesis itself were composed and originated entirely by myself while at ShapeSpace and in the Institute for Materials and Processes of the School of Engineering at The University of Edinburgh.

Esmé Anderson
Edinburgh, UK
May 2016

Acknowledgements

Thanks to Dr. Frank Mill, for teaching me how to research and for many bacon rolls;
to Dr. Andrew Sherlock for his industrial expertise and for ShapeSpace;
to Prof. Konstantin Kamenev for his support and guidance;
to my parents, my in-laws and my siblings; and to my wonderful man, Alan: **143**.
Mostly, thanks to the Creator and Sustainer of all things, the great I AM, without whom nothing is possible and in whom all things hold together.

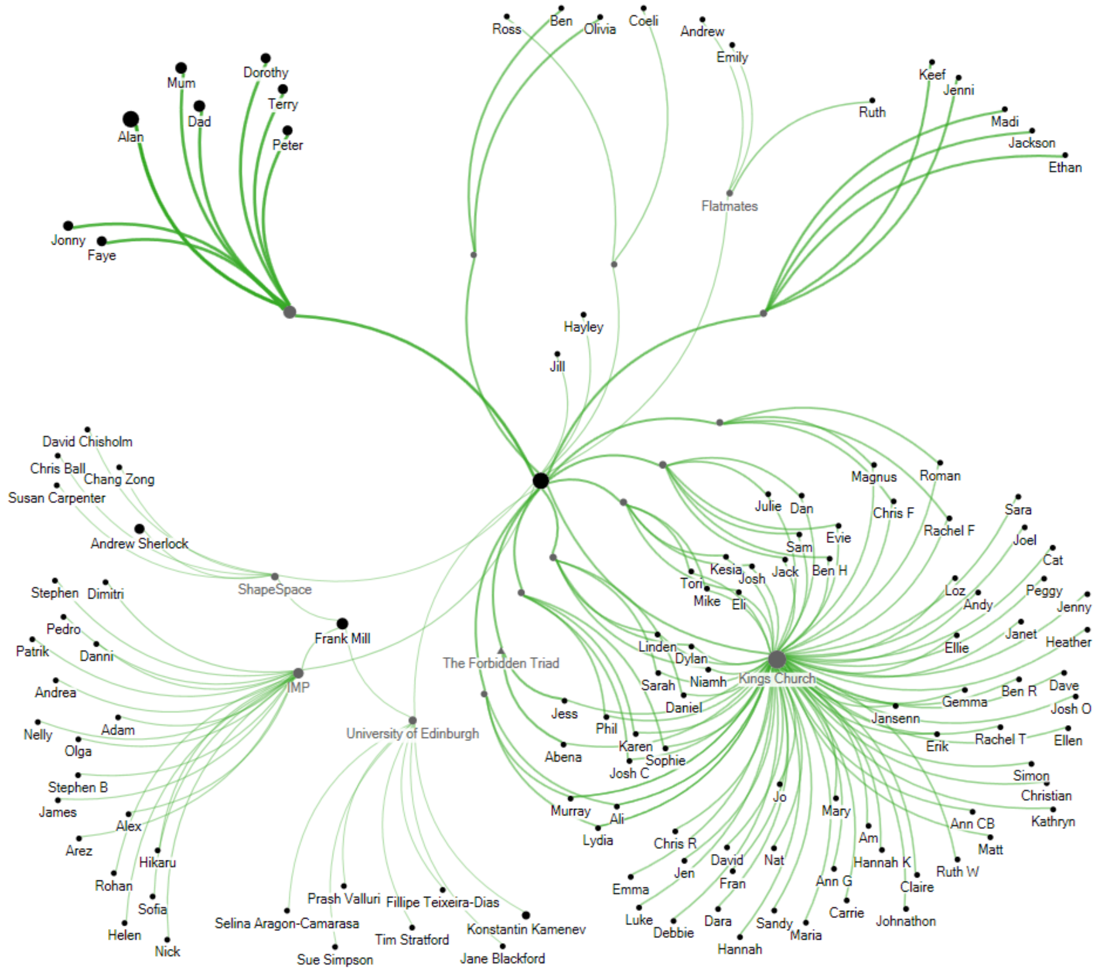


Figure 1: A thankful network

Contents

Declaration of Originality vii

Acknowledgements ix

Contents x

List of Figures xv

List of Tables xviii

Glossary xxi

1 Introduction **1**

1.1 Personal Statement 2

1.2 Research Rationale 3

1.3 Research Aim and Objectives 4

1.4 Research Contributions 5

1.5 Thesis Outline 6

2 Background and Motivation **11**

2.1 Initial motivation 11

2.2 Graph Theory 13

2.3 Network Theory 16

2.3.1 Six degrees of separation 17

2.3.2 The strength of weak ties 19

2.3.3 “Small Worlds” 19

2.3.4 Linked 21

2.3.5 Summary of network history 22

2.4 Network Visualisation 23

2.5 Computer Aided Design 26

2.5.1 3D shape search 27

2.5.2 Design reuse 32

2.5.3 Manufacturing and industry 36

2.5.4 Education and plagiarism 38

2.6 Summary 50

3 Methodology 53

3.1 Graph and Network Theory Basics 53

3.1.1 Nodes and edges 53

3.1.2 Graph metrics 54

3.1.3 Network images 56

3.2 NodeXL 56

3.3 Pajek 57

3.4 TinkerPop 57

3.5 ShapeSpace Technology 58

3.6 Edinburgh Benchmark 59

4 Investigating CAD networks 61

4.1 Introduction 62

4.1.1 Networks modelling CAD collections 63

4.2 Initial Mapping and Observations of Networks 1-5 65

4.3 Metrics and Measurements 72

4.4 Random Networks 75

4.5 Six Degrees of Separation 78

4.6 Small World Theory 81

4.7 Hubs and Connectors 83

4.8 Weak Ties 85

4.9 Layout 86

4.10 Discussion 89

4.10.1 Further work 90

5 Unimodal Networks 93

5.1 Introduction 94

5.1.1 Aims of unimodal network exploration 97

5.1.2 Proposed social dimension research 97

5.2 Network 6.1 99

5.2.1 General analysis method 99

5.2.2 Network 6.1 inspection 99

5.2.3 Network 6.1 results 107

5.2.4 Network 6.1 conclusions 110

5.3 Network 6.2 111

5.3.1 Defining plagiarism 112

5.3.2 How to create a cheat 114

5.3.3 Plagiarism detection method 116

5.3.4 Results 117

5.4 Networks 7 and 8 119

5.4.1 Cheats A7 and B7 120

5.4.2 Cheats X8, C8 and D8 121

5.5 Feature Analysis 124

5.6 Real World Application 128

5.7 Discussion 129

5.7.1 Critique of recent similar work 133

5.8 Conclusions 135

5.9 Further Recommendations 136

6 Multimodal Networks 139

6.1 Introduction 140

6.1.1 Aims of multimodal network exploration 141

6.2 Industrial Data Structure 141

6.2.1 Bill of Material data structure 143

6.3 Company A Data Structure Investigation: Network 9.1 145

6.3.1 Methodology 147

6.3.2 Initial results 149

6.3.3 Error in network creation 150

6.3.4 Correct network results 152

6.3.5 Further mapping: Network 9.2 153

6.3.6 Complete mapping: Network 9.3 156

6.3.7 Company A data structure discussion 158

6.4 Company B Data Structure Investigation: Networks 10.1 and 10.2 . . . 160

6.4.1 Initial mapping: Network 10.1 161

6.4.2 Further mapping: Network 10.2 164

6.4.3 Further visualisation of network 10.2 168

6.4.4 Company B data structure discussion 172

6.5 Data Visualisation 173

6.6 Discussion 180

6.6.1 Critique of recent similar work 181

6.7 Conclusions 181

6.8 Further Recommendations 183

7 Conclusions 185

7.1 Review of Aims and Summary of Thesis 187

7.2 Limitations and Scope for Further Research 189

A Publications 193

B Originality report 207

C Data Visualisation 209

Bibliography 213

List of Figures

1 A thankful network ix

2.1 The bridges of Königsberg graph 14

2.2 Fully traversable ‘house’ graph 15

2.3 Examples of Euler’s graph and a traversable graph 15

2.4 Snapdragon by Alison Turnbell, 2012 24

3.1 A basic directed network structure 54

3.2 Network of relationships in TinkerPop 58

4.1 Network 1: Initial mapping 66

4.2 Network 1: Adjusted visualisation 66

4.3 Network 2: Valve assembly in Sugiyama layout 67

4.4 Network 2: Valve assembly arranged by hand 67

4.5 Network 3.1: Edinburgh Benchmark parts modelled with assembly relationships. N.B. This is referred to as network 3.1 as later in this chapter different visualisation of the data will be presented, called 3.2 and 3.3 68

4.6 Network 4: Students’ CAD data linked by shape similarity. 69

4.7 Network 5: Large industrial CAD collection 70

4.8 Network 3.2 76

4.9 Network 3.3 76

4.10 Network 3.2 in forced circle layout 77

4.11 Network 3.3 in forced circle layout 77

4.12 Network 4 with clustering coefficient mapped to node colour 82

4.13 Network 4: Harel-Koren Fast Multiscale layout 87

4.14 Network 4: Fruchterman-Reingold layout 87

4.15 Network 4: Circle layout 87

4.16 Network 4: Spiral layout 87

4.17 Network 4: Grid layout 88

4.18 Network 4: Sugiyama layout 88

5.1 Network 6.1 produced from one class's CAD course submissions 100

5.2 Network 6.1 with duplicate edges merged 100

5.3 Network 6.1 with initial mapping of metrics to visual properties 101

5.4 Network 6.1 once 'ignore outliers' function has been activated 102

5.5 Network 6.1 with first visualisation of metrics 103

5.6 Example of alternative layout for network 6.1 104

5.7 Network 6.1 laid out by hand 105

5.8 Network 6.1 with further visualisation of metrics 105

5.9 Self-loops from network 6.1 shown in isolation 106

5.10 Final visualisation of network 6.1 107

5.11 Network 6.2 112

5.12 Predicted network feature highlighting cheat A6.2 115

5.13 Predicted network feature highlighting cheat B6.2 116

5.14 Network 6.2 including cheat A6.2 117

5.15 Network 6.2 including cheat B6.2 118

5.16 Network 7 including cheat A7 120

5.17 Network 7 including cheat B7 121

5.18 Network 8 including cheat C8 123

5.19 Network 8 including cheat D8 124

5.20 Graph of the Edinburgh Benchmark using feature relationships 125

5.21 Feature analysis network of cheat A6.2 125

5.22 Separated feature analysis network of cheat A6.2 126

5.23 Feature analysis network associated with student 049 127

5.24 Shape similarity analysis network associated with student 049 127

5.25 Network made to test plagiarism in a current class 128

5.26 Network 8 including cheat C8 with alternative visual properties 131

6.1 Expected overview network structure of industrial CAD Data 143

6.2 Graphic representation of a standard BOM data structure 144

6.3 Network 9.1: Initial network for company A 147

6.4 Expected plot shapes of company A's data 148

6.5 Plots of initial results from company A's data 149

6.6 Example of correctly built network graph from table 6.1 151

6.7 Example of incorrectly built network graph from table 6.1 152

6.8 Correct result plots from company A's data 153

6.9 Network 9.2: Partial model of the data structure of company A 155

6.10 Network 9.3: Complete model of the data structure of company A . . . 157

6.11 Network 10.1: Initial network for company B 162

6.12 Network 10.1: Initial network for company B with quantities 162

6.13 Network 10.2: Initial mapping of CAD item data 164

6.14 Network 10.2: Purchase order values mapped to edge widths 166

6.15 Network 10.2: Edge width representing edge weight within CAD data . 167

6.16 Network 10.2 with mapped edge weights 169

6.17 Network 10.2 rearranged 170

6.18 Network 10.2 with raw components labelled 170

6.19 Network 10.2 with edge colour applied 171

6.20 Network 10.2 mapped with many metrics 172

6.21 Network 9.2 in Fruchterman-Reingold layout 175

6.22 Network 9.2 in circular layout 176

6.23 Network 9.2 in grid layout 177

6.24 Network 9.2 in Harel-Koren Fast Multiscale layout 177

6.25 Network 9.2 in Sugiyama layout 178

C.1 What Makes a Good Visualisation? 212

List of Tables

3.1 Graph metric explanations 55

4.1 Basic metrics of each network 73

4.2 Classic interpretation of graph metrics and novel interpretation with respect to a CAD collection network 74

4.3 Metrics of networks 3.2 and 3.3 77

4.4 Clustering coefficient metrics for each network 81

4.5 Degree metrics for each network 83

6.1 Example of BOM data list format 151

6.2 Metrics of network 9.2 155

6.3 Metrics of network 9.3 157

6.4 Metrics of network 10.1 163

6.5 Key nodes in network 10.2 165

6.6 Metrics of network 10.2 168

Glossary

BOM	Bill of Materials
CAD	Computer Aided Design
ERP	Enterprise Resource Planning
MD5	A message digest algorithm
PCA	Principal Component Analysis
PDM	Product Data Management
PLM	Product Lifecycle Management

Chapter 1

Introduction

> **“...here is a secret that never makes the headlines: We have taken apart the universe and have no idea how to put it back together.” Barabási**

Network theory has become ubiquitous in daily modern life. Networks are now so commonplace that they are barely noticed as they surround us, connecting computers through cables, mobile phones through wireless technology, and people socially, who have become adept at using them often without realising; we have never been better connected.

The science of networks has allowed network theory to be advanced and used, not only in social science, but in computing, biology, chemistry and the physical sciences too. In his 2002 book “Linked: The New Science of Networks” Barabási documented the history and advances of network theory and considered the notion that everything is connected to everything else, suggesting that networks hold they key to understanding many varied areas and concepts [12].

In mechanical engineering, network theory is yet to be widely accepted or applied. Little work has made use of the progress reportedly made in other sciences. This work sets out to explore and analyse the uses of network theory within mechanical engineering, specifically in the area of 3D CAD design. There is great potential for network theory to impact this field, revealing connections and structure previously unseen, bringing insight and advancing understanding of 3D designs. Therefore this work seeks to explore the uses of network theory with relation to mechanical engineering design, using network theory to explore Computer Aided Design (CAD) data structures, beginning by looking at a single CAD assembly and expanding to consider a part library containing over 13000 models. This investigation also seeks to use this now popular science with another novel technology, shape based reasoning, using the results of geometric analysis and continuing that of Mill [14, 47, 131, 132] in exploring the uses of network theory.

Network theory can be understood to mean all the tools and developments associated with the use of networks. In this graphs play a central role because they can model any interlinking data structure, including specific properties about the data items and relationships. In this way network analysis is unique, providing a framework to analyse aggregate design data. Visualisations and metrics can be produced describing the shape and topology of networks giving unique insight about the modelled data. Network diagrams also provide a highly comprehensible way to view collections of information, and in leveraging the power of human vision may be much more effective than simple data tables.

This chapter begins with a short personal statement, followed by the rationale for this work. The aims and objectives of this research will then be presented, followed by a report of the unique research contributions contained in this thesis. This chapter will conclude with an outline of the thesis.

1.1 Personal Statement

The work here undertaken is the original concept of combining network analysis with advanced 3D CAD, proposed by Mill. This unique approach takes analysis most commonly applied to social network structures and explores its uses when modelling CAD data sets from different scenarios, testing the theory of whether things really are better when connected.

Throughout this work many new concepts have been learnt, many new skills gained and many new ideas understood. As with all research this work has occasionally been as simple as connecting A and B to C, while at other times it has had more in common with a space mission astronaut, motivated by the vague hope that there is something out there. This research has found successes, as well as limitations, through many existing, new and modelled relationships. The nature of networks means more and more connections have been seen, and more and more links made, in and through this work, convincing the author that everything *is* better when connected.

1.2 Research Rationale

During the 20th century graphs and networks were widely used to represent many areas of interest in mathematics, the sciences and the social sciences. Since then the world has embraced networks; the World Wide Web has expanded exponentially and network use in everyday life has become the norm. This includes regular use of computer networks within offices, frequent socio-economic network meetings for business people and excessive use of social networks by the public. Network theory is now widely used within science also and has been successfully applied to and helped solve problems involved in characterising networks of particles, molecules, computers and most commonly, people[12]. In 1965 Stanley Milgram published his famous, though frequently misunderstood study relating to small world networks and the so called six degrees of separation [130]. Interestingly, the purpose of his study related to finding a target node in a network, given a starting node, and was therefore a search based problem. The average degrees of separation, closely related to the average geodesic distance has become of considerable interest in many searchable networks, most obviously the World Wide Web. More recently many researchers have generated graphs and used them to analyse many modern web based social networks, e.g., Twitter or Facebook [80].

The advances made by network science in other areas have yet to be explored in mechanical engineering design. Therefore this work seeks to explore network theory in this key area. Manufacturing companies design and make millions of different products the world over, using CAD as a primary design method [40]. CAD models from commercial manufacturing companies are often considered the most valuable 3D data [47], and these are designed using 3D modelling software and Product Lifecycle Management (PLM), Product Data Management (PDM), or Enterprise Resource Planning (ERP) systems. Design engineers are working within the constraints of this environment and are often encouraged to recreate rather than reuse 3D parts, resulting in a loss of engineering knowledge, as well as time and incurring unnecessary expense.

This common problem has resulted in a very well-researched area: 3D shape search and matching. There are various methods and solutions and the topic has reached such a position that there is now a myriad of papers detailing ever more effective and novel ways of identifying shape similarity. However most of these are pitched at an abstract theoretical level and focus on accuracy of retrieval rather than on practical applications and solutions for design engineers. It is possible that the design problem facing engineers is larger than just reuse, and that the PLM systems they work with are limiting them because the real situation with CAD models is often more complex than these simply structured and constrained environments reflect. It may in fact be more akin to the less defined structures such as those found commonly in social networks and this is where this research begins.

Alongside this, as the government is aiming to grow the number of people attending university, many are training to become engineers and universities have seen more and more design students come to learn. A common and rising problem, with increased student numbers, is plagiarism of work. This is an unfortunate truth, but is a modern challenge that is faced by educators. With the proliferation of information freely available online set only to increase, action must be taken to ensure that the quality of students work is not compromised and academic integrity is maintained. There is currently a very effective and widely accepted software package used for the detection of plagiarism in writing, namely turnitin, which is used across all UK universities. However at present there is no such software available for 3D submissions. This research will therefore also undertake an investigation into whether graph and network theory alongside 3D shape recognition software can be used to help identify plagiarised CAD designs.

1.3 Research Aim and Objectives

The central aim of this work is to explore the uses of network theory when applied to collections of 3D CAD models and its uses when combined with shape similarity techniques, in a mechanical engineering context. This will be supported by investigating two main areas where CAD models are found, namely in manufacturing and education, both key areas of mechanical engineering.

The research objectives identified to achieve the aim are:

- Produce networks of parts collections from different scenarios, using shape similarity and other linking techniques to develop networks of CAD models. Use network metrics to measure the created networks and identify from network theory the most useful metrics for use with CAD part networks.

- Classify the use of network theory for analysis, search and visualisation of the created networks.

- Produce a network of a real world parts collection from the manufacturing industry. From this, collect information about the uses of network theory in this context.
- Develop the uses of networks in an educational setting, by investigating a class's design submissions, focussing on the role of networks and shape similarity techniques to assess the similarity of students' work, and thus develop a way to identify plagiarism within a class's submissions.

An in-depth review of associated literature will be conducted to assist in achieving this aim, including current 3D shape similarity techniques and network theory, alongside an exploration of the main developments in network theory. Both these areas have seen major progress and success. Building on the achievements in these areas, this thesis will focus on investigating the advances network theory can bring to mechanical engineering design, unifying this with the advances in shape similarity technology and being among the first to explore the uses of networks in mechanical Computer Aided Design.

1.4 Research Contributions

This work presents unique CAD networks; models of CAD collections linked via assembly structure or geometric similarity to expose the underlying data structures. Networks modelling mechanical engineering design data, corresponding diagrams and metrics have not been seen before, and this style of analysis has not been presented outside of Mill's research group. This work models collections of real world anonymised CAD data, from industrial companies and the University of Edinburgh, data sets that are uniquely available to this research.

Beginning this exploration of network theory applied to CAD collections this thesis presents five networks in chapter 4 with their associated network diagrams and metrics. These five networks are unique and the images produced display this data in a novel way. Their associated metrics are presented and discussed, and results suggest that these measures are significant in analysing CAD as individual assemblies and large part collections in education and industry.

Chapter 5 investigates unimodal CAD networks, modelling student submissions linked by geometric similarity, an entirely novel concept. A novel and robust method for detecting plagiarism in 3D work is presented and tested multiple times on sample data sets, before being trialled on a real world data set. The method is successful and shown to be effective, contributing to knowledge the ability to identify problematic submissions from engineering students.

In chapter 6 network models of large, real world data sets from mechanical engineering industry are presented. These multimodal networks include small overview structures, as well as partial and whole data structure models. The networks and their diagrams present the company's CAD collections and associated data in a manner not used before. It is proposed that these may be useful in assisting search and design reuse methods. This investigation received industrial feedback and used anonymised real world company data.

Also this work contributes unique diagrams of CAD collections, produced throughout the investigations using network theory, that have not been presented outwith Mill's research group. It is proposed that these images are effective in communicating information about the CAD data they display, contributing novel methods to exhibit mechanical engineering designs.

1.5 Thesis Outline

This thesis will now present an investigation into network theory and CAD collections. Chapter 2 presents relevant background material, first discussing network theory and the advances of network science as the initial motivation for this work. Graph theory, the foundation of network science, is presented in section 2.2 and network theory is discussed, with key advances reviewed in section 2.3. The field of network visualisation is also presented in section 2.4. Following this, Computer Aided Design (CAD) is discussed and the well-researched area of 3D shape search presented in section 2.5, before the issues of design reuse are introduced. CAD in industry and education will be explored in subsections 2.5.3 and 2.5.4, and the issue of plagiarism is discussed. The chapter concludes, having presented literature from different disciplines and branches of science, with clear observations, providing the starting point for this research, to equip the reader with a full understanding before the technical chapters.

Chapter 3 will introduce some of the methods used throughout this work. Specific methods are presented in the relevant technical chapters, however this chapter begins with the basics of graph and network theory in section 3.1 and presents the software used; NodeXL, Pajek and TinkerPop (sections 3.2, 3.3 and 3.4). Section 3.5 will give details of ShapeSpace's technology, before the Edinburgh Benchmark is discussed in section 3.6.

The first technical chapter of this thesis, chapter 4, pilots this investigation where network theory was used to model, analyse and provide novel information about various CAD collections. Five networks are presented, with relevant diagrams and metrics, and are analysed according to the key advances of network theory discussed in section 2.3. Network 1 models a simple CAD assembly and network 2 models two assemblies that share components. Network 3 models the Edinburgh Benchmark and is presented in 3 different ways, noted as network 3.1, 3.2 and 3.3. Network 4 seeks to advance shape similarity assessment, utilising these already accurate results to build a network of CAD files linked by geometric likeness, while network 5 models a large collection of over 13000 industrial CAD files. The findings of this initial investigation highlight the importance of unimodal, bimodal and multimodal networks, and CAD collections in industry and education. These results were further developed in two key areas, reported in chapters 5 and 6.

Unimodal networks of CAD data from education are explored in depth in chapter 5, with a focus on the uses of shape similarity results. Networks are built, showing the geometric similarity of 3D file submissions of students from the University of Edinburgh. Network 6.1 is created from historical CAD data and is initially analysed in section 5.2, with a detailed inspection presented in subsection 5.2.2 and subsection 5.2.4 concludes that this method may be useful for detecting unoriginal work. Section 5.3 furthers this investigation, introducing network 6.2 and defining plagiarism in 3D CAD work in subsection 5.3.1. Subsection 5.3.2 details the invented cheats A6.2 and B6.2, which are made from the available data. The method, adapted in order to detect plagiarism, is then detailed (subsection 5.3.3) and results reported in subsection 5.3.4. The method is simulated again in section 5.4, which discusses networks 7 and 8 with their corresponding cheats, and additional feature analysis discussed in section 5.5. The method is tested on a real world class's submissions (section 5.6) and the results are discussed in section 5.7. This chapter reaches some conclusions based on the investigation (section 5.8) and makes recommendations for further work in section 5.9.

Multimodal networks of CAD data from industry are investigated in chapter 6, using real world CAD collections and other associated data from companies A and B. The typical industrial data structure is predicted in section 6.2 and the structure of a Bill of Materials (BOM) is discussed in subsection 6.2.1, giving an understanding of a mechanical CAD design. The data from company A is investigated in section 6.3 and used to create network 9.1, 9.2 and 9.3; a methodology for analysing company A's data using network theory and initial results are presented (subsections 6.3.1 and 6.3.2). During this investigation a fundamental processing error was found. Issues arising are discussed in subsection 6.3.3 and a corrected network and analysis is presented in subsection 6.3.4. Further mapping of company A's data is presented in subsection 6.3.5, where network 9.2, a partial model of the data structure of company A, is shown. Subsection 6.3.6 advances this investigation, presenting a full network model of company A's data. This chapter then analyses company B's data in a similar way (section 6.4), beginning with an initial overview mapping, network 10.1 in subsection 6.4.1 and expanding this structure, creating network 10.2 in subsection 6.4.2. Additional visualisations of network 10.2 are presented (subsection 6.4.3), where the importance of metrics mapped to visual properties is identified before the investigation of company B's data is discussed in subsection 6.4.4. Section 2.4 assesses the visualisations of company data and critiques the available layouts for network diagrams, concluding that network diagrams may be useful communication devices for CAD data. This chapter then discusses the findings in section 6.6, drawing conclusions in section 6.7 and suggesting further work areas in section 6.8.

This work is concluded in chapter 7, where the results are summarised and reviewed with respect to the aim and corresponding objectives, outlined in section 1.3. The limitations of the work are discussed and suggestions for further work are made in section 7.2

The academic papers published through the course of this research are included in the appendices, along with a short background on data visualisation, relevant to work presented in section 6.5.

Finally, the reader should be aware of a few general points about this thesis. A point on a graph will be referred to as a 'node' and a line on a graph will be an 'edge'. A 'node' may be referred to as a 'vertex' in other work, however, as 'vertex' has several interpretations it will not be used in this thesis. This work presents ten original networks of different CAD collections, but it should be noted that some networks are presented in different views, or analysed in alternate ways. This is denoted through numerical sub-labelling, e.g.: networks 3.1, 3.2 and 3.3. One focus of this work is network diagrams, therefore throughout this work many images will be presented. These have been formatted for optimised viewing in print, including enhanced labelling and call-outs and may differ from those used in publications and other presentations.

Chapter 2

Background and Motivation

This chapter provides the supporting background materials for the following technical chapters as well as the motivation for this work. It is split into six sections, beginning with a presentation of the initial observations that inspired this work in section 2.1. In order to fully outline this work's background section 2.2 will consider the history of graph theory and section 2.3 will expand into the history of network theory and section 2.4 will discuss data visualisation. From there section 2.5 will introduce Computer Aided Design (CAD) as a key technique for mechanical engineering design with subsection 2.5.1 introducing the theme of shape based reasoning which, while it also finds its roots in graph theory, is a primary incitement for this work. Section 2.5.2 will discuss design reuse issues facing the industry while subsection 2.5.3 will explore other challenges being faced in the manufacturing industry currently. Section 2.5.4 will conclude this chapter, analysing in depth prevalent educational issues which precede those faced in industry.

2.1 Initial motivation

Networks have become commonplace in modern life. So widespread in fact that most people not only understand the basics of what a network is, but are adept at using them and do so daily. This has not long been the case however and the relatively quick growth and uptake of network technology has sparked the interest of many scientists and researchers. It is this that has inspired and motivated this work.

Within the past 40 years there has been a significant shift in the public understanding of networks; what they are and what the term means. Popularly understood as a social activity and associated with canapés, polite conversation and 'getting ahead', there is now a very different perception of the term.

Searching for the term ‘network’ using Google returns nearly 2.5 billion hits including varied explanations, images and even several films. Images consisting of people, computers, or dots connected by lines or arrows across a background of a map are common among the returned results. In the 1970s a film was released called “Network”, telling the story of a fictional television network. It can be said that this was the popular understanding of the term at the time. By contrast in 2010 a film called The Social Network was premièred, a retelling of the creation of Facebook, which signifies the change in modern understanding of the word.

The most noteworthy network can be said to be the Internet. After digital computers were invented in the 1940s, the World Wide Web was created by Tim Berners-Lee in 1989 [21]. Since then it has grown at an unprecedented rate. There were estimated to be over 3.5 billion users, just over 40 percent of the world’s population as of December 2014 [78]. This is a fascinating illustration of how one network has impacted and changed daily life.

Facebook is another household name that highlights this shift in public knowledge. It is a social media tool, which at its core depends upon and exploits all the benefits networks offer. Since its launch in 2004 it has grown to have 1.23 billion active monthly users. Over half of these have more than 200 ‘friends’ (i.e. linked to 200 other users), showing how powerful networks can be in connecting people. When it was first floated on the stock market in May 2012 shares were priced at £24. Sedghi reported the stock was worth £39 in 2014, and now the stock is worth over £82.57 (checked on 10/05/2016) a share, demonstrating that networks are not only powerful tools but can also be the basis of a profitable business [164].

It’s all thanks to computers that networks are so popular. Without this incredible technology, these networks would not have been created. As networking is commonly put to use in social situations it is not surprising that a science has grown around them, creating a prominent topic of study. Social science, concerned with society and relationships, often makes use of networks and their associated tools to analyse the structure and gain understanding of people’s connections. Networks are now also used in other sciences (see section 2.3) as their uses have become better known in the academic community.

Looking at networks and their development highlights how well connected the modern world is. Networks can show us that there are links everywhere, between people, organisations and countries, socially and professionally, at work and at home. With all these links it is unusual to find something that feels unconnected. However this work will focus on areas in mechanical engineering that are currently untouched by the advances that networks have brought to other sciences and will seek to assess whether they can be used to further these areas linked to mechanical engineering design.

2.2 Graph Theory

To understand networks it is important to first look at graph theory. Graph theory forms the very foundation of networks and as such helps us understand them more completely [13]. It is to be noted that in discussing graphs here we are not referring to the commonly misnamed 'charts' (graphical representation of data in lines, bars, or histogram plots with grids or coordinates) but instead we are referring to a representation of points that are connected via links.

Graph theory has its roots in Euler's now famous paper of 1736 [26, 58], though not identified by name, about the Seven Bridges of Königsburg. It was not until 1878, however, that Sylvester [180] introduced the term 'graph' and much later with König's 1936 textbook [107], "Theory of Finite and Infinite Graphs", that the term 'graph theory' gained scientific recognition. Since then, graph theory has developed as a major mathematical topic and has been extensively used in many areas, ranging from computing science, chemistry and physics to biology, linguistics and sociology.

At its core graph theory is purely mathematical, so it is surprising to learn that it grew out of a simple desire to solve puzzles, not complex, intellectual puzzles but everyday problems that were frequently discussed and well known. Perhaps less surprising is that these puzzles intrigued mathematicians and in solving them, graph theory emerged. The history of graph theory is explored in "Graph Theory 1736-1936" where Biggs *et al.* cover in depth the origins of graph theory and describe how it is now of use to mathematicians, chemists and physicists.

Euler's famous article on the seven bridges of Königsberg [58] was the first to address one of these entertaining puzzles. In it Euler presented a solution not only to the particular puzzle, but to all similar problems [141]. In 18th century Königsberg (now Kaliningrad) there were seven bridges spanning the River Pregel (now called Pregolya) and the residents are said to have spent their Sunday afternoons walking about the city [84], trying to devise a way to cross all seven bridges. In Euler's paper he addressed this problem, stipulating that each bridge should be crossed only once and in doing so created the diagram shown in figure 2.1.

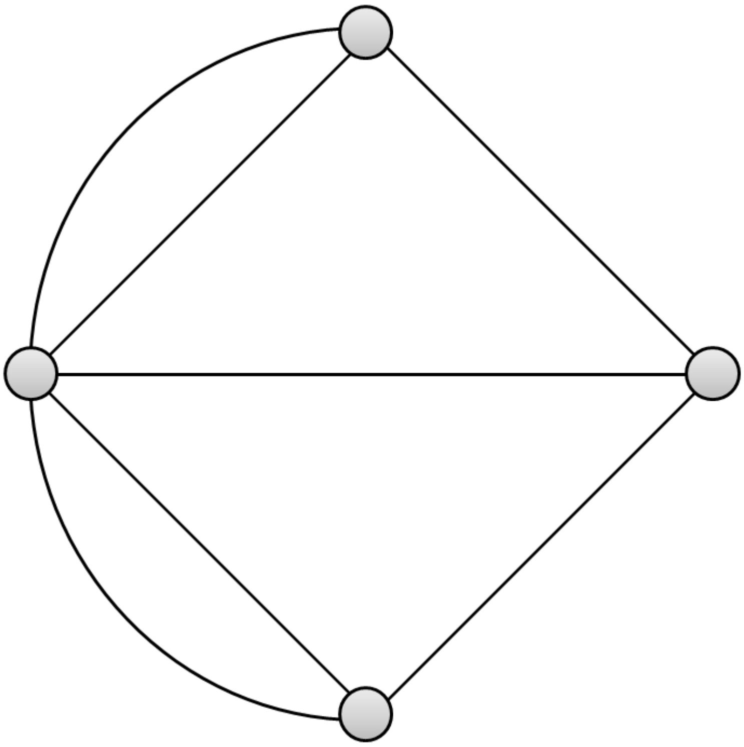


Figure 2.1: The bridges of Königsberg graph

He simplified the problem so each land mass or island became a point and the bridges became lines. The nodes (points) and edges (lines) created a graph shown in figure 2.1 and from this graph theory was born.

By considering how, starting at one node and ending at any other, each edge could be traversed (travelled along) only once Euler found that each node must have an equal number of edges leading to/from it for the graph to be fully traversable. In continuing this work Euler proved that for a graph to have this property it must have precisely two or no nodes with odd degrees (number of edges attached to it). This is well illustrated by a simple challenge, often posed to children during school. The challenge is to draw a basic house, shown on the left of figure 2.2 without lifting pen from paper. This can be achieved various ways, due to the inherent nature of fully traversable graphs, one of which is illustrated on the right-hand side of figure 2.2.

We see this is possible as the 'house' graph has two nodes with odd degrees and three

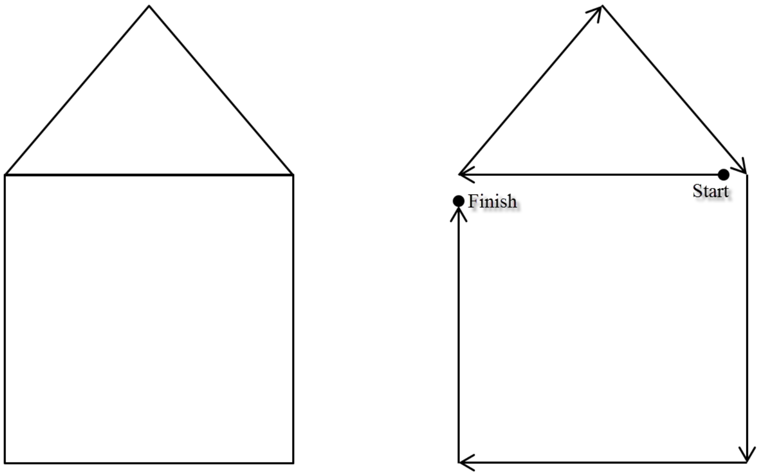


Figure 2.2: Fully traversable 'house' graph

nodes with even degrees. Another example to illustrate this is the image in figure 2.3, created by rearranging Euler's first graph. The first image shown in figure 2.3 cannot be redrawn without lifting the pen from the paper, but the second can. The degree of each node is shown, illustrating Euler's findings.

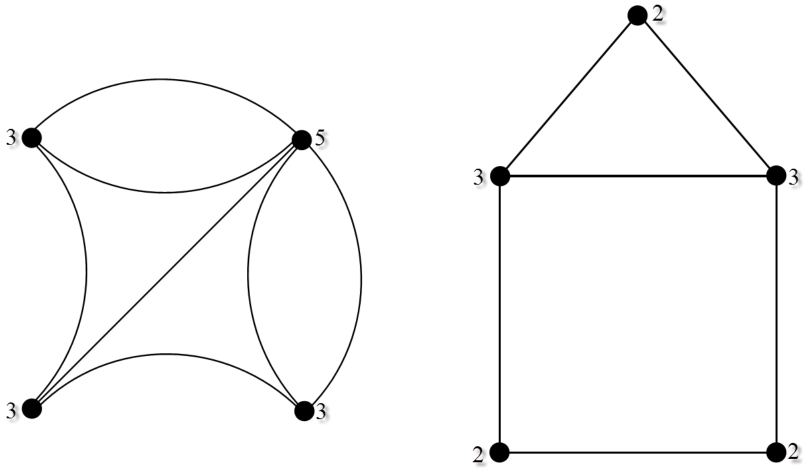


Figure 2.3: Examples of Euler's graph and a traversable graph

These simple examples show how Euler proved to the citizens of Königsberg that they would never be able to spend a Sunday afternoon walking around the city and crossing all seven bridges only once. Sadly few of the original seven bridges of Königsberg still exist (thought several have been rebuilt) [183] but thankfully Euler's proof has survived and is now often cited as the first true proof in the theory of networks [140].

Following Euler's article much work was done to advance the field by various scientists and mathematicians, including Kirkman who used graph theory to describe polyhedra [105] and Cayley who first explored trees [43]. Subsequently other scientists began to recognise the benefits of graph theory for their fields. Frankland was amongst the first to discuss graph notation in chemistry [66] while others, including Cayley [42] and Sylvester [180], also contributed.

After this the theory of networks became a popular subject of study and this will be discussed further in 2.3, where their history and advances in the field will be considered.

2.3 Network Theory

Networks are now a popular topic, with many applications in academia, and the science of networks is said to be the science of the real world [12]. There has been extensive research, the most famous work published by Milgram [130, 191] and Granovetter [77]. There have been notable summaries written by Barabási [12], Watts [194], and Buchanan [37] which this brief history and subsequent discussion are based on. An interesting discussion has also been written by Tesson (2006) in her PhD thesis, where she explores the history and development of network theory, and considers the conventional interpretation in relation to her posited idea that human communities can be treated as biological organisms [185].

Before considering the history of networks it is important to clearly define them and outline how a network is different from a graph. The extensive use of the term and its employment across differing branches of science makes this particularly complicated [194]. Many systems can be described using networks, so it is imperative to remember the data's context; a social network and a computer network are semantically different, but both are distinct networks [194]. The creators of Pajek (software that was utilised during this work and discussed in section 3.3) helpfully define a network in its most simple form as a graph which includes additional information on the nodes or edges [54]. Also Watts defines a network as "A collection of objects connected to each other in some fashion." [196]. In this work, this is the assumed definition of the term 'network'.

Perhaps fittingly, the development of network theory did not follow a neat, linear pattern. It emerged in the 1930s and was used primarily as a social science tool, with Moreno writing several papers exploring how networks could be used to analyse social groups amongst humans [136, 137]. The theory continued to be explored around the world by different research groups and cliqués [148] until the 1950s when Cartwright and Harary were the first to connect network theory with graph theory and mathematics [41, 199]. In their paper they explored the uses of networks in psychology, creating network graphs to advance Heider's theory of balance [81]. Cartwright and Harary are widely acknowledged to be the first to make the shift from describing networks in qualitative to quantitative terms, thus linking graph theory and network theory [199]. With this in mind a critical review of the most prominent publications in the area follows.

2.3.1 Six degrees of separation

In "The Small-World Problem" by Milgram (1967) the link between any two persons chosen at random was explored [130]. Milgram did this by choosing a target person, in this case a stockbroker who worked in Boston and lived in Massachusetts, and sent out letters to a random group of 'starting' persons within a specific city with exact instructions on how to forward the mail. The parameters were that participants should not send the mail directly to the stockbroker, unless they actually knew him. Instead they had to forward the post to a person they knew, who they determined had the best chance of knowing him. That recipient should do likewise, until the letter reached the target person. Of the 160 letters that Milgram sent out, 42 reached the businessman, in an median of 5 steps (the smallest chain was 2 and the largest chain was 10). From this, Milgram seems to conclude that an average of 5 intermediaries is needed to link any two randomly chosen individuals, while noting that it was most likely that a female would forward the mail to another female, a male to another male, and that people will more likely forward mail to friends than family.

This idea was perpetuated by many other researchers and notably by the playwright John Guare, who in his work "Six degrees of separation", posited that Milgram's theory would apply outside America to the whole of the world [79]. Milgram's other attempts to verify these results by repeating similar experiments received similarly low completion rates, however the Internet has since provided researchers with more tools and allowed further exploration of these ideas.

Inspired by Milgram, in 1994 Fass *et al.* created the Kevin Bacon game, in which every actor is linked to Kevin Bacon through cast lists, in as few steps as possible [156]. The game is still popular and has shown that the average number of steps between any actor and Kevin Bacon is in fact 6, meaning the game is often called the ‘Six Degrees of Kevin Bacon’. This shows how Milgram’s concept has been adopted into modern day life.

In his book “Linked”, Barabási questions Milgram’s measure and writes of experiments performed to determine whether it holds true in other networks. Continuing the line of investigation, Barabási set out to determine the degrees of separation on the Web. In 1998 the Web was modelled as a connected network was estimated to have around 800 million nodes and the experiments concluded the diameter of the web was 18.59 [3], otherwise expressed as 19 degrees of separation. Barabási goes on to report this has been determined as true in other areas, stating that the molecules in cells are separated by only 3 chemical reactions and authors from differing fields of science are 4 to 6 collaborations apart, showing many systems are more closely linked that previously thought. In fact the Web with 19 degrees of separation is reported as the largest [12].

In academia Watts conducted further experiments during 2001, where he attempted to once again to verify this result. His experiment involved emailing 4800 people with similar content as Milgram, this time with 19 targets. He found that the average number of intermediary steps between the first senders and the targets was 6 [196]. Microsoft also claimed to have proved this measure to be correct, as the Guardian announced in 2008 [174]. The theory was again confirmed in 2009 by Zhang *et al.* who performed further experimentation and mathematical analysis reportedly showing that the theory applied to on-line societies whose connection pattern may not be identical to real world social connections [207]. In 2010 Findlay reported that Facebook, which had 4.5 million users at the time, had a degree of 5.73 [64], another result which supports the 6 degree of separation theory. It wasn’t until Watts’ work on “Small worlds” became popular that this measure was accepted as common fact of network theory, see subsection 2.3.3.

In 2016 Edunov *et al.*, researchers at Facebook, attempted once again to measure how connected everyone in the world is. They reported that “Each person in the world ... is connected to every other person by an average of three and a half other people” [165]. Their work was widely reported in the media [34, 89, 189] as having empirically derived this value, however, rather than measuring the vast network at their disposal, statistical algorithms, in particular the Flajolet-Martin algorithm [65], were used to estimate this result. Their reported numbers, of an average distance of 4.57, corresponding to 3.57 intermediaries, are yet to be substantiated by other published work.

2.3.2 The strength of weak ties

Milgram's work on network theory was an important first step and in 1973 Granovetter published his key paper "The Strength of Weak Ties". In this he surveyed people who had been successful in applying for jobs in the Boston area of the United States. He asked whether they had been informed about the position by 'a friend' and found that often it was merely through an acquaintance that the job opportunity had been heard about. From this Granovetter posited that while networks commonly dealt with strong ties and their significance, it was in fact the weaker, more tenuous ties which played significant roles in social networks [77]. He concluded that weak ties between people are the most useful, providing them with new information and opportunities in ways that strong ties didn't. Granovetter was also the first to talk about network bridges; a link in a network connecting two closely linked clusters of nodes to each other. This is exactly how weak ties influence a network.

Friedkin tested Granovetter's theory of weak ties in more depth just one year later, by collecting information about colleagues within the same university [75]. After modelling a network of co-workers and investigating the strength of the ties between the participating academics, Friedkin concluded that Granovetter's approach did not resolve all associated issues and that links among groups of sub-area specialists in biological sciences display disproportionately weak ties [67]. Friedkin also stated that interpreting the significance of weak ties and bridges must be done carefully, as he was concerned that the existence of weak ties between social groups did not necessarily mean that important information would be transmitted along them.

Despite Friedkin's concerns, Granovetter's insights meant that networks continued to be investigated, with work being published in many varied disciplines. Networks were used to model systems including neural networks [1, 49, 83], oscillators [32, 112, 179, 201], arrays of electrical components [31, 200], excitable media [72], genetic networks [102] and the spread of disease [82, 90, 109, 121].

2.3.3 "Small Worlds"

Following this increase in the use of networks to model and analyse different systems the most significant advance in network theory was made by Watts and Strogatz in 1998. Until this point it can be said that networks were thought of as either completely regular or completely random. In their paper Watts and Strogatz explored network models between those two extremes, which more accurately reflected the biological, technological and social systems that networks were being used to model. Watts and Strogatz discovered that this midpoint was characterised by networks containing clusters of regularly connected nodes as well as some small path lengths, as found in random networks; in between regularly and randomly connected networks lay networks that contained elements of both. They termed these type of networks 'Small World' networks. Before this, models of disease spread indicated that the network structure and topology influenced the speed and extent of the transmission [82, 90, 109, 121], however Watts and Strogatz also reported that the dynamics of such a spread was explicitly a function of the structure of the network and not at all linked to its topology.

Watts and Strogatz proved that small world networks were neither random nor regularly organised, lying instead somewhere between these two extremes. They proposed that while their model had received little attention, it would soon be accepted as the most accurate descriptor of all levels of systems. They were correct in their prediction, and each year that followed this publication small worlds were found to effectively model systems in language [91], the World Wide Web [3], cell metabolism [99] and collaborative research networks in science [139]. Telesford *et al.* (2011) contested Watts and Strogatz's technique, claiming their proposed method, that identified networks as 'Small World' when they displayed clustering similar to a lattice and path lengths similar to a random network, was biased. Telesford *et al.* stated that this resulted in nearly all networks being classified as 'Small World', even those with very low clustering [184]. Instead they proposed a 'small-world metric' which they reported accurately identified this critical network type.

Watts has also done key work in the area of idea propagation and he is well known for discussing the concept of how ideas and influence spread from person to person. There are several conflicting views about how this happens amongst social scientists [75, 195], as well as disagreements over the ability and predictability of how individuals within a network can impact and influence other persons and the whole of the social network. Conflicting models have been built and arguments made based on these. Watts' (2007) computer simulations concluded that influence was spread gradually over time, by a critical mass of easily influenced individuals, rather than important singular persons, while Gladwell (2006) argued for a "Tipping point" that represented a sudden and dramatic change in a system. "Propagation" in a network can refer to ideas or trends spreading in a social scenario or can model for virus spread, however the concept is only relevant when the data being modelled by a network includes people and their relationships, where choices made determine the move of information, rather than a natural biological phenomenon. As such it is unlikely to be relevant to the research in this thesis.

2.3.4 Linked

After Watts the most popular and notable work published was “Linked” by Barabási (2002) [12]. Written not just for the academic community but also for a lay audience, this popular science book explains the depths of networks and the science surrounding them in beautiful, accessible simplicity. In “Linked” Barabási stated that nothing happens in isolation and “everything is linked to everything”. Following this principle he continues to experiment and advance the field. One focus of “Linked” is developing the work done by Milgram. Barabási and his colleagues tested the popular concept of ‘six degrees of freedom’ by investigating the diameter of the World Wide Web, concluding its size to be 19 [3]. As discussed in subsection 2.3.1, this theory is now widely accepted.

Barabási is also credited with highlighting the significance of hubs in networks. Hubs are nodes which have more edges connecting them to other nodes than would be expected within a network. The number of edges connected to a node is a metric called the ‘degree’. Barabási suggested that hubs are the most significant nodes in a network, as if they are removed they affect the whole network (and its metrics) far more than removing nodes with a lesser degree. From this Barabási introduced the concept of scale free networks, where the distribution of node degrees follows a power law. Scale free networks are still widely debated, but notably Watts disagreed with the theory that hubs are pivotal for a network as his 2001 experiments showed few paths included the crucial ‘hub’ nodes that Barabási’s work expected [196]. This debated area of network theory is unlikely to be useful to this work and therefore will only briefly be discussed in subsection 2.3.5.

Following Barabási’s work in this area, two other books of significance have been published more recently; “Small Worlds” by Buchanan (2002) [37] and “Six Degrees: The Science of a Connected Age” by Watts (2004) [194]. In his book Watts shared further insights about his theories in an approachable way, building on his already published work and making them accessible to the public. As he explored the science of networks he showed how they can be used to model diseases and Internet viruses, the 1977 New York power cut, and the success of Harry Potter [194]. Buchanan clearly outlined and discussed Watts' small world theory and showed how it can be applied to many and varied situations in life, including food webs for ecosystems, proteins in yeast, and the Internet [37].

2.3.5 Summary of network history

This short network history illustrates how, from humble mathematical beginnings, graphs and networks have been developed into a key academic field. Much research has been done and is still being undertaken, advancing this area, and finding uses for and validating network theory. It is clear that this cumulative research has opened up many more possibilities for network theory to be investigated.

After graph theory was established, network theory was investigated by Milgram, who is credited with discovering the theory of six degrees of separation. Subsequently Granovetter explored the theory underpinning the strength of weak ties, concluding that weak links were more important than strong links in social networks. Watts and Strogatz then investigated the nature of networks, discovering that most are neither entirely ordered nor entirely random, instead lying somewhere in between, and dubbed these 'small world' networks. Following these, key advancements have been made by Barabási, Watts and Buchanan, whose work continues to be at the forefront of this field.

Many of these developments have been undertaken in a social science setting, where of course research has been heavily biased towards social networks. As network theory has advanced however, it has also been utilised in other sciences, including chemistry and biology.

This huge area of research has had several clear outcomes. Firstly networks were once considered static, fixed data sets but as understanding has advanced, this has been shown not to be the case. Watts stated that "Real networks represent populations of individual components that are actually doing something" [196]. Networks have been shown to effectively model many different types of dynamic, fluid systems that are continually changing, evident in the wide array of scenarios reported in the literature, and this justifies further research. Secondly network theory, while well established, still contains many areas of disagreement. These areas, including the importance of hubs and small world theory, warrant further investigation and are likely to be useful in the future.

As this area has become well recognised in the academic community it is worth noting that network theory is also widely acknowledged and used by the public, and this accounts for a few of the references in this section which refer to newspaper articles, plays, films and Internet publications. In particular, the Guardian's article on Microsoft's research shows that networks have become commonly understood and used, and are seen as valuable in the public's eyes [174].

This short overview of the history of network theory shows that huge advances have been made in the field. Of the main ideas in network theory, three areas, namely idea propagation, scale free networks and power laws have only been briefly mentioned as they are outside the scope of this research. While useful in many contexts, they will not be examined here. For further discussion on this see section 7.2. As research has continued network theory has been proved to be applicable almost anywhere, to any system or collection. Comparatively little work, however, has been applied using network theory to any area of mechanical engineering design or shape search. Mitchell (2006), considering the work of Barabási and Watts, deduces that the general science of networks is relevant for search of all kinds [133]. It is with this in mind that the research in this thesis is undertaken.

2.4 Network Visualisation

Networks are readily transformed into images, which are of great worth when it comes to effectively communicating data. Visual inspection of network diagrams can reveal details about data that may not otherwise be evident, making them high-value images. In this section network diagrams, their attributes and layouts will be discussed while other types of data visualisation are considered in Appendix C.

Data visualisation is a broad field, and draws on many areas including graphic design, advertising and visual arts. There are many advantages to using images to communicate and the popular idiom "a picture is worth a thousand words" rings especially true when data is combined with images to communicate and engage. Network visualisation is a powerful tool for communicating information as it naturally combines data and image.

As network theory has developed so has network visualisation. Becker *et al.* (1995) claim there is a rich history of network visualisation due to the high value of network data [20] while Shneigerman (2006) highlights how successful visualisation can communicate meaningful, high level information from a simple overview [168]. Bertin was the first to explore network visualisation in 1981. He introduced the idea of manipulating how a network was displayed, interactively. Bertin also discussed the stages of decision making, stating that data and statistics are not enough to reach well informed conclusions, but visualisations of data are key in helping this process [22].

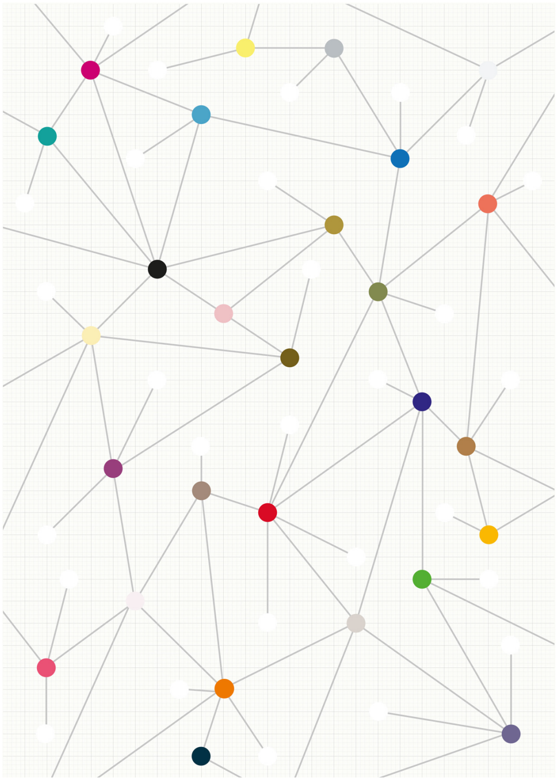


Figure 2.4: Snapdragon by Alison Turnbell, 2012

Subsequently, work has centred on utilising the existing techniques and the optimisation of images. While researchers have continued this work, others, including artists, have exploited networks in their work. Alison Turnbell is once such artist, who takes graphical representation of data, such as diagrams, blueprints and plots, and transforms them into abstract art. Figure 2.4 shows a popular piece of her work, entitled Snapdragon (2012) which was inspired by a 19th century scientific attempt to classify colours [187].

Network visualisation has received a huge amount of academic attention, warranting entire conferences devoted to the topic. In 2014 the "International Symposium on Graph Drawing", which had been running for more than twenty years, extended its name to the "International Symposium on Graph Drawing and Network Visualisation", to acknowledge the importance of the area. More recently the University of Edinburgh began their own conference "Dealing with Data" which has a strong emphasis on data visualisation techniques.

There are many graphical features, including shape, width, length, symbols and colours that can be used to add visual impact and clarity to a network diagram. While a network is simply made of points (nodes) and connecting lines (edges), it is possible to use graphic tools to enhance and change the appearance of the diagram and so focus attention on key content and relationships it portrays. Network visualisation can be split into the following categories:

- **Layout** The layout of a network diagram can be changed manually or by using several developed layouts.
- **Labels** Labels can be applied to nodes or edges, or groups within a network diagram, to exactly communicate some information.
- **Directed edges** Directed edges have arrows attached to one or both ends to represent the direction of the links
- **Node attributes** This data can be shown using variations in shape, colour and size of the node.
- **Edge attributes** This data can be shown by variation in colour, curve and width of the edge.

In an earlier paper Becker *et al.* stated that these visualisation possibilities, particularly for node attributes, can have a great effect on the interpretability of the resulting image [19]. Clearly these attributes should not be decided upon without knowledge and consideration, lest a network diagram be presented poorly and therefore interpreted incorrectly.

Other notable work done in the area of network visualisation includes that by Sarkar and Brown, who created a new way to visualise network graphs using an interactive fish eye lens [160]. Also Paulisch discussed network layout algorithms and their constraints, as well as focussing on particular parts of the layout, and attempting to introduce the idea of a collapsed node, representing a group [143].

All these developments led to a vast array of analysis tools and software packages being developed allowing researchers to explore networks, each with varied methods of visualising the data input to them. In 2011 Cobo *et al.* undertook a review of 9 different mapping tools, comparing the different analysis techniques and tools available to researchers and concluded that visualisation can help to interpret and analyse results, stating “data need to be visualised for us to get to grips with it” [48]. This work will make use of several of these tools, which are presented in chapter 3.

In 2013 Huang *et al.* produced a visualisation of 40000 student code submissions [88]. This image was among the first of its kind and Huang used it to compare different ways of using code to solve the same problem. Comparing students’ work in non-text submissions is complex, but networks could provide a platform to identify similarities between designs within the mechanical engineering industry and education. This will also be explored in the scope of this thesis.

2.5 Computer Aided Design

Mechanical engineers design and make millions of different products the world over. There has been a notable rise in the number of Computer Aided Design (CAD) models generated, as they have become integral in product design, engineering and the manufacturing industry [40]. Many of these models belong to companies, who employ PLM (Product Life cycle Management), PDM (Product Data Management) or ERP (Enterprise Resource Planning) systems to manage their design and manufacturing processes. Each 3D CAD model contains valuable geometric information which is stored in the management system. Design engineers work within the constraints of this environment and often recreate rather than reuse 3D parts, resulting in wasted time and money and a failure to make the best use of engineering knowledge, encapsulated in the CAD software and models.

CAD files from commercial manufacturing companies are often considered the most valuable 3D data [47] and many researchers agree that CAD is a high value topic [85, 110, 204]. 3D models are created by skilled design engineers, who are trained and experienced in modelling and specialist software, arguably making these practitioners highly valuable personnel. Due to the intrinsic value of 3D model data, problems identified in engineering design have become leading research areas. Design reuse is one such commonly discussed problem with many possible solutions, including the prominent and well-researched area of 3D shape search and matching. There are various solutions as is evident from the numerous papers detailing ever more effective and novel ways of identifying shape similarity. Many of these focus on accuracy of retrieval rather than practical applications within industry and solutions to solve the problems faced by design engineers.

Shape search and matching are most often attempted by comparing and measuring the similarity of two CAD models. There are many different methods of comparing shapes, which are discussed in subsection 2.5.1. The proliferation of methods has not yet resulted in the industry finding a reliable way to begin to improve design reuse through shape search techniques. Some of these methods have been used to create software aimed at helping product development teams find CAD models [29], but there are few results reported. While there are other design reuse options, focussing on shape statistics results seems a promising direction of research that could be furthered using network theory.

2.5.1 3D shape search

The proliferation of CAD models and the discussion of design reuse for engineers has led to the well-established field of 3D shape search being developed. Effective summaries have been written on techniques, algorithms and retrieval methods by Cardone *et al.* (2003), Bustos *et al.* (2005), Iyer *et al.* (2005) and Tangelder and Veltkamp (2007). Effective shape search necessitates effective shape similarity methods where geometric shapes can be compared, therefore enabling search by shape.

Shape similarity work has focused on matching techniques and search capabilities, where varied and multiple methods have been developed to return ever more accurate results. Much work done on shape similarity is focused on industrial CAD and retrieval methods, though some work assesses CAD in other situations. There is an intriguing congruity in the aims of shape search technology and the network Milgram (1967) first investigated to search for a target person [130], as Milgram's aim to 'find' the stockbroker is similar to an engineer's search for a particular part in a CAD collection. Also it is interesting to note that graph theory, discussed in section 2.2, is probably the birth place of shape search, as from this polyhedral graphs and topology were developed. The connections, between sections 2.2 and 2.3 and 3D shape search, highlight the nature of networks to find and investigate linking relationships, and so may be useful in exploring further ways these areas could be used together or linked.

Assessing the similarity of two shapes is a geometric problem which, thanks largely to computer science applications, has become a well-researched area. There are multiple ways to tackle the problem, and many researchers have done so, resulting in a wealth of successful approaches. Within these varying approaches, it is possible to choose different parameters to measure the similarity between shapes. It is worth noting that despite the vast array of possible parameters, those chosen are often a biased reflection of the desired outcome of the researchers. Due to the large amount of work in the field, the accumulated wealth of knowledge can seem like an academic problem solving game, where ever more accurate methods are pursued, but with little real world testing to justify the practical value of the results.

There are two main steps for any method of shape similarity comparison. The first is to compute a shape signature for every item in a collection and the second is to compare all these using a suitable function [40]. Despite the variation in methods, all research conducted must find a way to compare shape signatures two at a time, in this case two CAD models. This is done by generating a shape signature for each item analysed, denoted as S(x) for model x. There are many different ways to generate this signature so this is a key area of investigation. The similarity between two shapes or models can then be expressed as a similarity metric [95] or as the distance between the two related shape signatures. This must be done many times over to assess a sizeable collection of 3D CAD models. Santini and Jain (1996) were among the first to discuss similarity matching and they presented equations for measuring similarity [159]. These measures have since been refined and within the field most papers agree shape signatures of items x and y respectively [S(x), S(y)] should have the following properties, where δ(S(x), S(y)) denotes the distance between the two shape signatures:

- **Positivity**: There must be no negative measures of the distance (difference) between the shape signatures, as negative values of similarity have no meaning:
  $\delta(\mathrm{S}(x),\ \mathrm{S}(y)) \geq 0$

- **Identity (Self-Similarity)**: If two shape signatures are the same, they should have a distance measure of zero to show they are geometrically identical
  $\delta(\mathrm{S}(x),\ \mathrm{S}(y)) = 0 \therefore x = y$

- **Symmetry**: The distance between the two shape signatures should be identical regardless of the order of comparison. The measure should be the same whether x is compared to y, or y is compared to x. This measure means distance is not affected by the order of comparison
  $\delta(\mathrm{S}(x),\mathrm{S}(y)) = \delta(\mathrm{S}(y),\mathrm{S}(x))$

- **Triangle Inequality**: In some applications triangle inequality must be satisfied by the distances between three shape signatures, x, y and z, and is best described

by the following:
$\delta(\mathrm{S}(x),\mathrm{S}(y)) + \delta(\mathrm{S}(y),\mathrm{S}(z)) \geq \delta(\mathrm{S}(x),\mathrm{S}(z))$

- **Invariances**: The measures should be independent of rotation and other translations. They should be invariant with respect to the underlying representation of the models. Chaouch and Verroust-Blondet (2009) present a novel method of aligning 3D models before they are compared, as an alternative to other normalisation steps in shape comparison methods [44].

- **Robustness and sensitivity**: For any shape, any magnitude of change in the shape signature should accurately reflect the magnitude of change in the actual shape, object or model. If this is not adhered to a small change in $\mathrm{S}(x)$ may result in $\delta(\mathrm{S}(x), \mathrm{S}(y))$ being disproportionately changed with δx. A small change may lead to two similar objects being measured as markedly dissimilar. Poor sensitivity in this area will result in major errors.

- **Computational efficiency**: However the shape signatures are compared, it must be efficient, as the aim of the comparison is to quickly and reliably assess large collections of shapes, objects or models. If the time taken for comparison is relatively long, the method will not stand up to comparing many parts within a large collection [40, 182].

Shape signatures are compared to each other to fulfil selected research goals and each variation in how shape comparison is determined has its own strengths and weaknesses. Some methods centre upon comparing features and others prioritise the comparison of certain measures, while all are affected by the type of shape signature they generate. This has resulted in many shape similarity methods being developed and these can be divided into several categories. They can simply be split into feature based methods, graph based methods and others, however Cardone (2003) identified seven main titles [40], providing a comprehensive set of groups, by which most major contributions can be categorised:

- **Features**: In this method shape signatures are evaluated based on type, size, orientation and number of features they contain and on their interactions. This method doesn't consider the gross shape of the object. Bustos *et al.* (2005) asserted that feature extraction was the initial most important development in 3D object retrieval [39].

Saupe and Vranic (2001) used feature vectors to compare 3D objects [161] and Jun *et al.* (2001) presented a method that compared feature shape descriptors using neural networks [101]. Chu and Hsu (2006) combined feature analysis with topological graph methods, but this method was unable to cope with simple modifications such as chamfers or fillets [45], while Wei and Yuanjun (2007) created feature vectors based on voxelisation of analysed models [198]. Leng and Xiong (2009) combined feature analysis with visual characteristics in their TUGE method, to improve upon retrieval [116]. Many researchers, including Bai *et al.* (2010) use feature analysis to attempt partial retrieval; Bai *et al.* used features to create an associated graph [10], Bronstein *et al.* used partial retrieval to compare a Centaur and a Horse [35], while Li and Godil (2010) used feature analysis methods to compare different retrieval methods [119]. Lmaati *et al.* made use of feature vectors and concluded this method to be robust for noise and decimation [120] and Philipp-Foliguet *et al.* (2011) used global and partial feature techniques to compare artwork 3D models [144].

- **Spatial Function**: Methods that fall into this category use shape signatures that are spatial functions, for example they may employ a Gaussian map that maps a set of points for a solid to a reference sphere.
  Reuter *et al.* (2006) introduced the idea of creating a shape-DNA, or 'fingerprint' for 2D and 3D manifolds [151] while Jain and Zhang (2007) presented a method that mapped shape descriptors onto an affinity matrix [96]. Key work has also used Principal Component Analysis (PCA) [175, 167] and Zheng *et al.* (2008) introduced a successful local scale-based model, focussing on retrieving mechanical parts [209].

- **Shape Histograms**: Sample points on the surface of the 3D models and representative characteristics are extracted in this method. Measured characteristics are organised in histograms, which are then compared. The efficiency of this method varies inversely with the number of sample points that are chosen.
  Rea *et al.* performed work in this area, combining histograms with neural networks (2004) [150] and proposed new shape distributions (2005) [149], as did Wang *et al.* (2008), who developed a new shape signature based on D2 shape distribution histograms [192]. Mademlis *et al.* (2009) introduced a novel approach to shape matching they called the 'Shape Impact Descriptor', where shape signatures were formed based on the model's impact on the surrounding space [124]. They as-

serted that the geometry-based retrieval results were good with their method, but it was of little use when trying to retrieve semantically similar models. Mademlis *et al.*'s approach is arguably an example of unnecessary development; though another efficient shape signature has been devised, it does not improve upon the uses of shape similarity or matching. Kuo and Cheng (2007) proposed another method, but this was determined to be too computationally complex for real-time application [111] and Itskovich and Tal (2011) used histograms, representing shape distribution index, to address partial matching in archaeology and suggested their method could effectively reduce the time-consuming nature of digital archiving currently practised [93].

- **Section images**: In this method sections of the object are taken and used as the shape signature. The sections are taken at various places and then inserted into a neural network. This classifies the signatures into groups, based on the group technology code used. This method is not invariant to scaling, translations or rotations but it is robust.

  Filaliansary *et al.* (2005) presented an image based method where 3D models were represented by characteristic views, even comparing photographs to their shape signatures and returning good matching results [63]. Similar methods were used by Wang *et al.* (2008) [193] and Li *et al.* (2010) who claimed their image-based method was superior to others [117]. Zhu *et al.* (2010) developed a method that mapped 3D models to 2D images and argued that it performed better than others [211].

- **Topological graphs**: This method simply compares topological graphs that are used as shape signatures. The graphs can carry extra information about the objects they represent on their nodes and edges.

  Baron *et al.* (1999) notably used voxel representation to optimise shape, a technique which would be used to optimise shape signature creation [14]. Iyer *et al.* (2005) and Siddiqi *et al.* (2007) presented methods that created skeleton graphs as shape signatures [94, 169] while Cheuk Yiu and Gupta (2007) used 3D scanners to produce 'point clouds' [92]. Biasotti *et al.* (2008) used topological Morse theory to define shape signatures [24] and size graphs [25], and Tierny *et al.* (2009) improved upon previous Reeb graph methods [188], while You and Tsai (2009) and Ma *et al.* (2009) used B-rep graphs [205, 123]. Ma *et al.* concluded that their method was effective for middle sized data sets, but impractical for real life or industrial solutions [123]. Larabi (2009) combined graphs and text descriptors to produce an XML language that would describe shape signatures [114]. Xiaoliang *et al.* (2010) also made use of B-rep graphs [202] as did Ma *et al.* (2010) in their investigation of automatic discovery of common design structures [122]. Zhu *et al.* (2012) used Laplacian spectrum graphs and show this approach disposes of the singularity problem often met when analysing CAD models [210].

- **Shape statistics**: Methods that fall into this category use a coarse comparison of basic geometric properties, such as volume and surface area, to measure similarity between models. This technique is quick and efficient but does not provide enough discrimination for accurate comparison.
  Kazhdan *et al.* (2003) used this technique to present a symmetry descriptor, where reflective invariance is exploited to identify geometrically symmetrical models [103]. Clark *et al.* (2006) used this method to compare shape signatures to human perceptions of similarity, concluding that the method accurately matched human perception [47].

Attempts have been made to classify CAD models in ways other than in the categories presented above. These include categorising 3D models by their appearance, their function or by the manufacturing process used to produce them. Kopena and Regli (2003) described the importance of linking engineering designs by function, using a 'semantic web' [108], following Szykman *et al.* (1999), who presented a schema for representing function [181]. Bustos *et al.* (2007), while reviewing content-based retrieval methods, stated that classifying shapes should ideally be automated in industry [38].

There has been work focused on partial retrieval or combining techniques; Hu *et al.* (2007) compared various techniques to investigate the influence of feedback on model retrieval and concluded that retrieval performance can be improved using parallel solutions to compare similarity [86] while Akgül *et al.* (2009) discussed the use of relevance feedback and a combination of shape descriptors to give better performance on 3D shape matching [2].

2.5.2 Design reuse

Design reuse is a key issue within CAD and engineering manufacture and the aim of much shape search work is to aid design reuse for industry, but using shape similarity methods to identify CAD parts and so inform reuse is not straightforward. In their review of CAD search techniques Bespalov *et al.* (2005) argued that most techniques perform poorly on real CAD objects [23]. They are not the only ones to have commented on the issues shape similarity techniques have with CAD models, however there have been advances since Bespalov *et al.* wrote their review, and Falcidieno and Herman (2011) highlighted the important role of shape matching in driving design reuse and also discussed the importance of semantic matching [59].

Design reuse is not limited to engineering design, but includes architecture and other design areas. In 1999, Sivaloganthan and Shahin wrote an overview of design reuse techniques, critiqued the methods and discussed many of the issues. Among their review of the relevant work they identified some key design reuse issues, arguing that design reuse can kill creativity in industry and so hinder, rather than aid, advancement. They also highlighted the issue of accurately recording problems, such as poor design or errors. They concluded that integration would be key in design reuse tools becoming successful and predicted that computerisation would be a way forward in this [172].

Xu *et al.* (2007) claimed to show that utilising product family structure was an effective strategy for encouraging re-use, however they asserted this from an acknowledged biased start point [203]. In 2009 Tomiyama *et al.* gave an overview of design methodologies and generalised that design theory was not widely taught, understood or applied in industry. They briefly commented on design reuse when discussing the issues facing design [190], but did not expand to discuss whether it would effectively aid designers. Ferreira *et al.* (2009) attempted to encourage reuse among engineering designers by building a shape 'thesaurus', where shapes could be compared to an existing database, but the method was limited as it had size restrictions [61].

You and Tsai (2009) outlined the importance of shape similarity assessment within the product design process, stating that reuse of designs, and therefore knowledge, from previous components increases a designer's efficiency [205]. Bai *et al.* (2010) developed a semantic based method for CAD retrieval, which they concluded was effective [10] as the approach developed was design reuse orientated. However, the increased accuracy of shape search techniques do not solve the engineering problems and Altfeld *et al.* (2011) commented on the lack of industrial evidence of design reuse advances being put into practice [7]. In 2012 Dongmin *et al.* proposed another reuse system for a manufacturing context, where companies were modifying their businesses from selling products to selling services, in a bid to increase revenues, and so require reuse capabilities to inform product design, use and maintenance. They developed a framework they claimed was effective, evaluated by interviewing experts from industry and academia, who shared their opinions on the framework, but the experts did not use it, nor was it implemented by a manufacturer.

After more work had been done, Silventoinen *et al.* (2014) reviewed available reuse techniques in relation to customer-oriented design and presented a comprehensive list of challenges, suggesting developments needed. They concluded that design reuse must be considered from many perspectives and organisational practices and IT systems should be used to support reuse. They also identified a lack of clarity about the effects of design reuse and whether it limits the creativity of designers.

Despite the differing techniques and varying success of methods, nearly all papers written on the topic agree that design reuse would significantly reduce developing cost and time in the mechanical engineering industry [198] but this issue of uptake is yet to be resolved. There has been a lack of work addressing this or focussing on how design reuse would be put into practice in an industrial situation. Also researchers often note that the complexity of their work makes it unrealistic for the real world [111], and comment on the problems caused by PLM systems.

Iyer *et al.* (2005) identified the need for shape search within engineering, highlighting that as new personnel are added to engineering design companies, previous search methods would have limited success. They also highlighted that PLM and other management systems, while linking all models with their associated keywords, may not help the process, due to the most valuable information about a CAD model being primarily geometry-related or dependent [95]. Falcidieno and Herman (2011) asserted the importance of keeping non-geometric information with graphic data [59] which is in part what PLM and other management systems do, but did not go on to discuss common problems with management systems as Iyer *et al.* did.

Despite much research focused on aiding design reuse, with limited real world success change in industry is reportedly slow. In 2014 Zhang *et al.* identified that commercial CAD systems (PLM, ERP, etc.) were widely acknowledged as ineffective in aiding designers in reuse [208] and Bai *et al.* agreed in 2016, stating that efficient design reuse, and thus knowledge reuse, is a key issue [11]. Despite many advances, the discussions in these recent papers show that the problem is far from resolved and shape similarity work is limited in its reach.

It is also notable that the design problem facing engineers may be larger than just reuse. The management systems used and customs among practising designers may be limiting them, as the real world situation with CAD models is often more complex than these simple-structured and constrained design environments reflect. Huang *et al.* (2015) agree that geometric reuse techniques are limited, as they lack contextual knowledge. In their 2015 paper they presented a method to reuse detailed design knowledge and discussed the issues facing design reuse, including the issue of repeating poor design through cursory reuse [87]. Their method does not provide a robust solution to the problems identified. Huang *et al.* also commented that, even though they had developed an effective reuse method, it would require extra time and commitment from the user [87]. They argued that additional work from current designers would save others time, however this is a key part of the issue with uptake and this highlights the deficiency of methods that aid designers without adding to their work. This is a prominent example of the disconnect between reuse methods and real world designers.

Demain and Fruchter (2006) highlighted that design reuse cannot occur solely through corporate memory (such as PLM/ ERP systems) but through social interactions between colleagues or with a mentor [55], proposing that design practitioners prefer to ask colleagues who have worked on similar projects and problems rather than use software. In their ethnographic study of design reuse, Demain and Fructer argue that design reuse happens through social knowledge networks and identified that these networks were created by social events within companies, where knowledge can flow, people are linked and knowledge seekers and providers are connected. While undertaking their study they focused on knowledge reuse, unlike other work discussed here, as they did not want to limit reuse to previously designed components. They conclude that social knowledge networks are crucial to reuse, however they acknowledge the network they identified is an informal one and not physically defined, freely using the term to categorise interactions between people. Since this work, little other work has focused on this network form, however there are clear links to the patterns of knowledge reuse identified by Demain and Fructer and the work on network theory presented in section 2.3. Mill (2013) has agreed with this assessment and noted that the design process itself, not just reuse, may in fact be more akin to the less defined structure found commonly in social networks [131].

The collections of CAD models that are available and are used in much of the research presented here and subsection 2.5.1 are benchmark collections or sets, developed for research purposes. The Princeton Benchmark collection and the McGill Shape Benchmark are two such collections that have been developed, and others have since been made available. The Princeton Benchmark was developed in 2004 at Princeton University and is a database that contained annotated 3D polygonal models. This database was created with an emphasis on shape matching, and the creators posited that the collection would be effective for shape matching methods [166]. The McGill Shape Benchmark was developed in 2005 at McGill University and prioritised models with articulating parts, again with a focus on search and retrieval [206]. There are many other shape benchmarks available, including those introduced by Jayanti *et al.*(2006) [97] and Fang *et al.* (2008) [60], but this work will use the Edinburgh Benchmark, developed by Mill [14, 47, 131, 132], presented in the methodology (section 3.6). This part collection has been built with an emphasis on engineering collections of CAD models, developed to accurately represent real world scenarios instead of focussing on shape search, matching or retrieval methods as have other collections. As such it is seen as preferable, as network theory in relation to CAD models is investigated, to remove the common biases work in shape searching has included by using the Edinburgh Benchmark collection.

This work will begin by assessing network theory methods in relation to this extensive data pool and proposes that there will be uses for network theory in mechanical engineering design environments. Rather than focussing on accuracy or improving shape similarity or design reuse techniques further, this work will employ the results of shape similarity analysis, using this as data with which to build a network. Shape similarity results provide robust and reliable data with which to investigate the uses of network techniques within mechanical engineering design. They form a sound basis for investigating the links between CAD files, which is a novel area for research. Network theory may reveal helpful insights for design reuse, as this work develops. It is key to this research that the context from which these CAD models come is understood and as such a brief overview of two key sectors in which CAD models are found will be discussed.

2.5.3 Manufacturing and industry

Design and manufacture within the mechanical engineering industry is not the only sector to use CAD as a primary design tool; architects, civil and electrical engineers, and a spectrum of product designers, from jewellery and textiles, to furniture and personal electronics, use CAD to realise designs. CAD, especially since the 50s has sped up the design process and impacts all levels of the manufacturing process [62]. In the manufacturing industry CAD is important as it allows ideas to be modelled and tested before production begins. As such, CAD has become the dominant method used by engineers, who are specially trained in these techniques. Xiaoliang *et al.* asserted this in 2010 and stated that 3D models are still the main mode of design and fabrication in the manufacturing industry [202]. Due to this there are large collections of current and historical data in manufacturing companies.

These sizeable collections of CAD data represent a significant amount of effort and expertise, as well as expense. They have inherent value to the company or designer who owns them, and are the most valuable 3D data around [47, 198] because they are linked directly to profits. CAD data also has value as models which accurately represent designs and also as key tools for communication, between persons working on all aspects of a product life cycle [110, 132].

Over the years, lean manufacture and processes throughout the product life cycle (supply and product processes) have been a key focus of industrial companies [170], reportedly successful in reducing wasteful practices, and have become accepted standard practice within industry. There is a disconnect, however, between now highly lean manufacturing processes and the design process. The majority of engineering companies employ PLM or similar systems to organise and manage their 3D data, but these systems are not effectively managing the design process and are comparably poor at streamlining design procedures. PLM systems serve a tightly structured purpose, outside of which they have been said to stifle the business and prove restrictive to design practice [138]. Rather than allowing companies to freely design in a lean fashion, PLM systems are considered to constrain and involve laborious procedures, which promote wasteful practice.

In an attempt to improve these practices, research into methods such as shape search and design reuse have become popular areas of study. As CAD parts are seen as valuable in industry, this research has received large amounts of funding, shown by the wealth of research discussed in subsections 2.5.1 and 2.5.2. Jones (2012) claimed all companies know they have a problem in this area and, focussing on the issues caused by duplicate parts, asserted the average cost to a company of having a single superfluous, duplicated CAD part was around $10,000 [100]. It has been noted that the design problems facing engineers are possibly larger than just reuse. Another recognised issue is part variety and there has been work done aimed at reducing the diversity of valuable parts [7].

ShapeSpace have worked with engineering companies that own huge quantities of design data, all producing cutting edge, innovative designs of cars, trains, aeroplanes and other mechanical products. There are also many engineering companies providing the industry with essential components such as valves, switches, and other hydraulic components. Through collaboration with ShapeSpace, this research will have access to real world CAD collections and parts, to which networks can be applied, so their uses in the real world may be assessed. Working with large industrial manufacturers, ShapeSpace have seen than many of them recognise the problems associated with current design practice and PLM systems, including CAD model proliferation, but to avoid public criticism there is little common acknowledgement. Also companies do not have time, knowledge or appropriate resources in house to deal with the problems they face.

The PLM systems engineering designers work within may be limiting them, due to the fact that the real situation involving CAD models is often more complex than these simply-structured and constrained environments reflect. It may be more akin to less defined structures, such as those found commonly in social networks. Models based on nature are useful as we look at the product life cycle [132] and as such there is an argument to be made for exploring the uses of networks, as they have been used to model other systems [197]. This work (chapter 4, Network 5) will begin looking at the possibilities network theory can provide in analysing large collections of industrial CAD parts.

Other work that has proposed methods for using networks in design industry includes that done by Li (2005), who presented the idea of collaborative design. However the research presented in this thesis focuses on using communication networks to allow collaboration [118], rather than utilising networks to analyse viable design options or address the design issues industry faces.

2.5.4 Education and plagiarism

CAD models from industrial manufacturers are known to be highly valuable [47, 198] and subsection 2.5.3 identified the large collections located within industry. Another notable area where sizeable quantities of 3D designs are found is within educational institutes, where students are trained to design and use specialist software. As CAD is seen as valuable within manufacturing, it is reasonable to say that education and training in CAD subjects is also valuable. This supposition is validated by the research that has been conducted in this field. Engineering designers, formerly referred to as draughtsman [62], need to be well educated and trained to become effective in industry and universities are key educational institutions where this happens.

In universities students are commonly taught CAD through a mixture of lectures and laboratories to train them in design and specialist software skills, in combination with education on other design techniques. Knowledge and understanding is then often assessed through written coursework and examinations. Some examples of CAD course methods could be termed 'solution design', where students are given a problem for which they must design a correct solution [76, 85, 110], or 'situation design', where students are given a brief they must design to resolve [5].

Some CAD education focuses on design skills, while other courses concentrate on equipping the student with knowledge of software. Design practice is key in this environment and much education is focused on students becoming competent and responsible designers. In Scotland there is a drive for education to highlight sustainability at school level, but there is little to suggest this is carried through to university education by set regulations or guidelines for educators. At the University of Edinburgh one of the main methods of CAD education involves a series of lectures and practical training sessions in labs where students are assisted in tutorials. These are aimed at teaching design techniques and encompassing specialised software training, before students are given a design brief assignment that must be completed individually. These types of courses are typically assessed via coursework submission of a written report and the 3D models the student has designed. Since 1994 CAD has been taught using Solid Edge, in the second and third years of the mechanical engineering by Mill, a senior lecturer and fellow of the IMechE, who has produced specialist training tutorials based on his expertise.

While discussing CAD in the automotive industry, Field (2004) highlighted the importance of education for designers. He highlighted how training and education for CAD continue to increase but also discussed the importance of university and other learning, stating that solid geometry should be taught to give engineering designer a good and firm foundation of understanding [62]. In 2004 Werner Dankworth *et al.* argued that how CAD is taught should be extended to cover not only design but all phases of the life cycle in a production chain. They proposed that design is a process and that instruction restricted to surface or solid modelling techniques was lacking. In their paper, Werner Dankworth *et al.* reported on education and training within industry and within universities, suggesting the limitations inherent in timetables at academic institutions may limit students' knowledge, and therefore their ability, in industrial situations [52]. Also Ye *et al.* (2004) highlighted the importance of CAD education, presenting a unique, industrial perspective, as they discussed the specifics of what should be included within CAD courses in universities. They evaluated the importance of students being correctly trained and prepared for industry, by collecting the opinions of prominent and experienced teachers [204]. In 2005 Piegl reviewed the ten most prevalent challenges in CAD, and included CAD education among them. In his report he chastised academic institutions and called upon them to train and educate students fully in CAD, which he claimed they were failing to do effectively. Piegl suggested comprehensive textbooks and a journal focussing on CAD education would improve the training of students as well as the state of CAD overall [145].

Alemzadeh and Burgess (2005) briefly reported on how CAD is taught as they presented a CAD project based at Bristol University, which they assessed as successful in equipping students with the core competencies in CAD principles and practice. They noted that in response to industrial changes, universities and colleges have professional CAD/CAM systems and focus on projects involving design and manufacture, where students are required to use CAD practices including geometric modelling [5]. Robertson (2007) examined creativity in relation to the use and education of CAD and argued that design instruction should highlight good engineering practice in and out of any specific class and train students to exercise good engineering judgement. Robertson reported that the focus on CAD in engineering teaching may unintentionally discourage creative problem solving [152]. Research into CAD education is continuing still, with Gracia-Ibàñez and Vergara (2016) most recently presenting a new method involving action research for CAD teaching, developing self-learning material and using rubrics for assessment. They reported that the practices were effective after observations over a two year period.

With much work done on advancing of CAD education, the issues prominent in the field are also widely discussed. Ye *et al.* (2004) assessed what should be included in CAD education using questionnaires to gather opinions on existing systems and agreed with Werner Dankworth *et al.* that it should focus on more than design alone [204]. Rossignac (2004) outlined the argument for education-driven research in CAD, concluding that several of his developed techniques were effective in teaching CAD [154]. He used education-driven research, distinct from Gracia-Ibàñez and Vergara's (2016) action research, to simplify more complex subject matter in CAD education and deduced that students would be able to understand core competencies and topics quickly, allowing them to advance, however this method was not assessed in any real world scenarios. Rossignac agreed with Piegl's assessment of CAD education, stating that engineering

CAD textbooks were rarely written, or presented as desirable accomplishments for educators, as universities do not focus on or encourage these types of achievements.

Another key issue within engineering education is plagiarism, a topic which has risen to the forefront of academic awareness. Plagiarism is not limited to engineering and as such much of the discussion of this issue has taken place in alternative subjects, including other sciences, humanities and social sciences. Plagiarism is a considerable threat to the integrity of academic work, and given that CAD education is linked to industry it is notable that academic integrity is implicitly linked to professional integrity [69]. A prominent divide in the discussion of plagiarism issues is between text based and non-text based work. Text based plagiarism has been a key issue in education for years, popular enough to gain attention in the press [50, 18] as well as much research from the academic community and even conferences dedicated to the subject. Popular and widely available software, such as turnitin, has made detection of problem students or unoriginal work possible, but has not solved the issues, as the vast quantity of recent work illustrates. While engineering education often makes use of written coursework and examinations to assess students' knowledge, it is unique in assessing learning via 3D CAD models and designs, however there has been little discussion of plagiarism directly linked to 3D, non-text based work. In order to most effectively discuss this issue, examples and work from within other sciences shall be discussed here, while some examples of non-science work will be included. Other subjects, such as visual arts, computer science and music courses, that are similar to CAD education in non-text based assessment are far removed from it in subject matter. It should be noted there has been much study done in the field of text-based plagiarism and thus this is where much of plagiarism theory is based.

Plagiarism has been defined in many ways by different researchers, all with differing emphasis. The Oxford Dictionary defines plagiarism as "The practice of taking someone else's work or ideas and passing them off as one's own" [56], the University of Edinburgh concur in their advice pages [186], and others have gone on to further refine this explanation based on their context. Culwin and Lancaster (2001) interpreted student plagiarism as "Plagiarism with the intent of gaining academic credit" [51] while Park (2004) suggested plagiarism in education is better termed "academic malpractice that should be deterred, detected and dealt with appropriately" [142]. As plagiarism has become a popular topic, there has been a lot of discussion about the ethics of cheating. Integrity and ethics have become commonly used words by educators. Striving to teach students integrity can feel like a battle and it is a common view amongst some educators that people are fundamentally liars, who lie in context but still think they are honest people, and that integrity is the exception to the norm [8, 69, 74]. This is not explicit within research, but can be inferred from the wealth of work done in this area.

Gabriel (2001) quoted Wilensky in a newspaper article, claiming students leave high school unprepared for the intellectual rigours of college writing [68], suggestive of the reasons why students turn to plagiarism in the first place. As much work in this area has continued, researchers from both text and non-text based subject areas have sought to identify why students plagiarise and how to stop it from occurring, though their methods vary.

In 2001 Lee compared web-based interactive manuals to traditional methods for instructing laboratory-based subjects and discussed how plagiarism was an inherent problem, as student collaboration outside of the laboratories made cheating more likely. While suggesting several methods with which to resolve these issues, including discussing and presenting results during the lab, Lee does not conclude these methods were effective in combating plagiarism despite presenting a novel instruction technique and these methods are not applicable to all non-text based subjects. Lee also claimed that the proliferation of personal computers was partly responsible for the rise of the plagiarism issue [115]. While others have agreed with Lee, there is little empirical evidence to support this position other than common observation [50, 57] and Gabriel (2001) reported that digital media was having an effect on student's academic integrity, as 40% admitted to plagiarising by copying a few sentences while 20%, down from 34% in the early 90s, would consider copying 'serious cheating' [68].

Allen (2003) stated that the Internet is a copy machine and mechanism [6], in agreement with Lee, while Piegl (2004) briefly highlighted the advantages and disadvantages the Internet would bring to CAD, including the possibility for a 'depot' to be created, where CAD models could be 'sold', and collaborative design possibilities, where work could be completed by partnership and sharing. He briefly reported on the issue of security, stating that it was the most important issue, as security, specifically the privacy of information, must be ensured if the other advantages he presented were to be exploited. Rossiter (2011) argued conversely that technology can enhance student learning. Piegl (2005) and Rossiter (2011) agreed that educational issues specifically within engineering and thus including CAD, are in part due to the tension between research and teaching within educational institutes [155]. In 2014 Sikanthi and Asmatulu (2004) agreed with Allen that copy and paste capabilities provided by the Internet would undermine academic integrity [177].

Gabriel (2001) reported on one case, where a student defended his decision to copy from Wikipedia, as it is counted as common knowledge, at the University of Maryland [68]. Notably Wikipedia's founders wish the site to be know as a reputable encyclopaedia,

> "Wikipedia's mission is this idea of imagining a world in which every person on the planet shares the sum of human knowledge. And that is what we're doing. But also, for free, and in the language of your choice"

Wikipedia wish to be a quotable, reliable source and so there is an unresolved tension while educational institutions are discouraging students from using it as a resource and the founders are aiming to become a legitimate source of reference for all areas of education [158]. In this case the student could have used the information there to inform the writing, but if the student were more informed, they may have been able to cite the work they were referencing. It is disputable whether the student's copying, the lack of referencing or the availability of the information on Wikipedia is the issue. Martin *et al.* (2006) reported, on text based work, on the effectiveness of the Joint Information Systems Committee's Plagiarism Advisory Service (JISC PAS) stating that students did not perceive it to be an effective system and allowing students to see the results of the comparison report produced heightened anxiety about being accused of plagiarism. Martin *et al.* stated their research was focused on making copying material an irrational choice for a struggling student. Although they recognise the JISC PAS could be part of a holistic approach to combat plagiarism, it does not provide a complete remedy. Their '3D' approach, based on deterring, detecting and dealing appropriately with plagiarism, as defined by Parks [142], does not focus on non-text work.

While Robertson (2007) did not mention plagiarism as an issue in CAD education, it can be seen that reduced emphasis on creativity, as he supposed, may cause students to be unaware of how to produce innovative solutions and designs, and therefore be tempted to turn to plagiarism as an easy solution.

Learning thought copying is also an issue in mechanical engineering design teaching. Students are often taught by creating 3D drawings from pre-existing technical drawings or downloading parts from websites such as Grab CAD. There are many parts available online, in a similar fashion to the 'depot' suggested by Piegl [145], but the inherent problem with this availability that is students may present this work as their own. It has been observed, though contact times spent in design laboratories with student at the University of Edinburgh, that they are concerned with presenting advanced, complete designs and this can overshadow the learning of specialist techniques, which is the aim of the class. Students often enjoy having a goal which produces something physical or measurable, rather than learning outcomes that result in them gaining knowledge and being trained comprehensively in specialist software. During these classes and their resulting assessments, if students download, use and reference parts from online sources, it is deemed an acceptable use of resources. Without referencing these stock parts, the students would be considered to have plagiarised. This behaviour becomes more complex, however, when students use these stock parts as inspiration or copy them to learn CAD modelling techniques.

Crace (2007) discussed the issue of when inspiration becomes plagiarism, in non-text based work, and cited famous examples of artists, such as Hirst, who has been involved in several law suits due to his work being plagiarised. Crace reported Blythman as commenting on how some forms of work do not allow for referencing, such as interior design, where houses do not contain a small plaque acknowledging influence, and commented that despite the lack of clarity in defining plagiarism, it was obviously important that something be done about the issue [50]. Economou (2011) considered that copying for inspiration becomes a problem when the inspirational material becomes the solution and the creative process is sidestepped. She highlighted how this circumventing puts into question the integrity of any such design process, or in this case assessment. Economou discussed the issue of copying from the Internet specifically with regard to graphic and digital design practices and outlined how creating documentation for educators and students provided a framework and lecturers to explain to students what plagiarism is and how it should be avoided through good practices. Garrett and Robinson (2012) presented a brief but comprehensive overview of the issues stimulated by discussing visual arts and non-text plagiarism [70]. They discussed how learning though copying was a key part of the creative process, agreeing with Economou and Blythman and also considered the issues of there being no clearly defined way for referencing or acknowledging visual or non-text material in a bibliographic sense.

Rossiter (2011) argued that technology should be used to enhance engineering education [155] and identified the advantages of allowing students to make audio recordings for learning and use of animations to enhance learning. Rossiter also briefly commented on plagiarism and suggested that the first year of university was the time to education students about what plagiarism is in order to prevent it from occurring.

As this issue has come to the forefront it has even been reported as an international issue, with the BBC reporting about students in India who argue that they have the right to cheat, unsurprisingly creating a knowledge and education gap between those who can afford to pay for answers and grades and those who can't [98]. Stappenbelt (2012) explored whether cultural differences between students, where they are originally from and if they study in the same place has any impact on plagiarism within engineering, in response to other work presented. Stappenbelt argued that cultural differences do have an impact on plagiarism and again concluded that educating students clearly about what plagiarism is and how to avoid it, discouraging students from resorting to copying, would be an effective way to prevent plagiarism [178]. Garret and Robinson (2012) agreed with Stappenbelt about cultural differences being influential in a student's perception of plagiarism and recommended that clearly established briefs for projects would be a helpful solution [70] as did Sikanth and Asmatulu (2014) [177].

Chuda and Naurat (2012) discussed software plagiarism and worked with students to report first-hand views on student plagiarism. Using similar methods to Ye (2004) [204] and Gracia-Ibàñez [76], they constructed a questionnaire to gauge student opinions and concluded that to work closely with students on such a sensitive topic was novel. They presented interesting statistics on student and staff views of plagiarism but did not make any recommendations on how to improve the situation [46].

Sikanth and Asmatulu (2014) examined the methods by which students cheat in engineering subjects, assessing the effects of modern technology on dishonesty in exams and other submitted work. They focused on text-based work in engineering, such as exams, and recommended, yet again, that educators discourage students from cheating through various methods, involving education [177]. They argued that engineering students would more commonly cheat, due to advanced technology being available, with male students cheating more than female students. As education becomes more high-tech so do the cheating methods, making them more difficult to detect and deal with. Sikanth and Asmatulu also evaluated the reasons students cheat and stated that social or family pressure and expectations, perceived disadvantages and competition contribute.

In 2015 Morales and Dominguez reported that final year students who were challenged on their plagiarism most commonly did not understand what plagiarism was and didn't understand the problem of copying or presenting unoriginal work as their own. This work, a decade after Martin *et al.* (2005) [125] and others suggested education was the key to stopping plagiarism, shows there has been little improvement in this area. Morales and Dominguez concluded, as much work previously had, that educating students about plagiarism is the most effective method of prevention. They stated that plagiarism should be considered as another educational tool in a students' experience and education [135], in contrast to Park's suggestion of deterring, detecting and dealing appropriately with plagiarism.

Unlike text based work, 3D models are not simple to retrieve [182] and this is true when considering plagiarism in education settings, as well as in industry. There are many resources and established pieces of software that educators can use to assess the similarity of text-based work, but few comparable tools are available for non-text based work. Google has developed image search software and provides the ability to search for images that have been reversed, and TinEye and iThenticate provide similar image search capabilities, however none of this software has provided a robust and reliable technique for detecting unoriginal work and it does not apply to 3D work. They are far from the efficacy of turnitin and have yet to be widely accepted. Arrish *et al.* (2014) assessed plagiarism of figures and charts within text based submissions and discussed a method for detecting plagiarism of 2D images and other figures using flowcharts. They commented on the lack of similar work to detect non-text plagiarism, despite the importance and serious nature of identifying unoriginal work [9], again indicating of the lack of methods to detect 3D plagiarism.

While software such as turnitin is commonly used to assess plagiarism in text based work, it does not provide an effective solution for detecting unoriginal work in mechanical engineering courses. The submissions for these assignments are rarely just text-based and students are expected to hand in models of their designs; single components, sub-assemblies and full assemblies are submitted via on-line forms or physical memory, but it is near impossible for a teacher or lecturer to assess the similarity of the many hundreds of models they are given. The research in this area has not produced a method for detecting plagiarism in 3D CAD based work and there is no current way to measure the similarity of students submissions in this subject area.

Iyer *et al.* (2005) highlighted the importance of defining similarity before attempting to measure it [95]. Investigations on the topic of shape similarity all define similarity and aim to achieve a successful way of measuring it. The summary of work done in subsection 2.5.1 shows how much success there has been in this area. However, it does not follow that the same measures of similarity can be used when looking at educational plagiarism.

Sanna *et al.* (2012) reported on their use of shape similarity methods in education, specifically 3D design, when they outlined a fully automated method for assessing 3D modelling exam submissions. Sanna *et al.* assessed an exam where students were required to design a solution, judged to be correct based on how similar their solution was to a given reference object. They propose a method, similar to shape similarity techniques, that measured the similarity of the students' solutions against the reference object and automatically produced a grade [157]. Sanna *et al.* reported on the efficiency of this method and asserted that students respected the approach, but made no links to the possibility of assessing plagiarism, or the inherent problems it would cause with their proposed method.

Houjou (2013) introduced an evaluation method for CAD submissions and asserted that 1/3 of CAD students in Oyama National College of Technology admitted to submitting material that was not originally theirs [85]. The method presented claimed to help students avoid plagiarism, which was achieved by providing students who were behind, and therefore likely to cheat, with a model solution from which to work. Houjou asserted that this method was effective in preventing plagiarism, however it failed to address the issue, evading the core problem and providing students with an easy answer which they didn't need specialist knowledge to produce. While this method would be effective in assisting struggling students, it may also inadvertently aid those who would be likely to cheat due to idleness [68] and thus fail to promote academic integrity.

Browning (2014) proposed that creativity and plagiarism are inversely related, so plagiarism could give students greater opportunity to produce creative and inventive designs. Conversely, Gino and Ariely (2012) demonstrated that creative people are more likely to cheat [74]. Browning argued, through an analysis of Bob Dylan's work, that collecting ideas and influences, and then presenting them as one's own is plagiarism because credit is not being given to the source, and this practice limits true creativity [36], while Gino and Ariely suggest that creative individuals have increased ability to justify their dishonest behaviour. It can been seen that a student who is heavily influenced by other's work will be more likely to submit unoriginal work, however Browning does not suggest a method to combat this problem and if Gino and Ariely are correct, creative students taking engineering design courses are likely to copy or plagiarise and rationalise their behaviour.

Gallent (2014) asserted that cheating is the norm and integrity is the exception among people, not just students and Keegan (2014) agreed, illustrating many popular and famous authors and poets have plagiarised not only ideas, but actual texts [104]. While some would argue that many cases of plagiarism are due to a lack of understanding, Gallent stated that only 15% of exposed cheats claim ignorance, with most declaring they cheated due to stress and pressure. Gallent argued that we must shift our perspective, accepting that plagiarism is going to occur and we should educate students about why academic integrity is important and also argued that academic integrity and professional integrity are equal, as a student who copies will become an employee, a politician or a designer who copies [69]. If it is the case that students must be educated about integrity, plagiarism detection software will not be an effective tool in reducing plagiarism and instead courses should be adapted to teach students how to maintain originality. Chuda *et al.* (2012) presented a unique insight into this when they asked students directly 'Is plagiarism wrong?' in their questionnaire. They found that only 30% responded that plagiarism was wrong, 1% said that it was not and 69% gave no response. They suggested that it was safe to assume that students had insufficient information to answer the question, but the work of Gallent, Keegan and others would indicate otherwise.

Referencing is another issue for academic integrity, with some claiming that citation is rarely taught correctly, compounded by a lack of citation standards for non-text based work [50, 70]. There is no accepted way to reference the original author of an influence or image that has been used when the new work does not include text. Porter (2014) commented on this, proposing that acknowledgement is key and insisted that IP rights must be observed. Porter commented on the use of TinEye and Google image search, both of which only address problems for 2D images [147]. Simon *et al.* also commented on the lack of a standard referencing system within non-text work and examined whether it would be effective, demonstrating that there are blurred lines between plagiarism and traditional practices for learning. Simon *et al.*, building on the work done by Blythman *et al.*, suggested that interviews with or presentations by the students could help assess the students' understanding of the work they had submitted [27, 171]. Within CAD education this could be effective, but not in all non-text based work; in coding education and practices there are often times a student may not be able to clearly explain exactly why code works and runs correctly though it is their own work, but copying another's code may give the same student a better understanding of of how and why that code runs.

In contrast to written text, it is unlikely that a student will have the ability to quote or reference another designer in their 3D modelling work. Having access to industrial CAD models would mean that students could be assessed on where they have derived their inspiration, but as discussed copying can be a valid form of learning in education, especially in the context of students being taught to design using specific software. That being said students must not be trained to simply mimic previous or current popular designs, but it is arguably a valid learning technique and it may be undesirable to prevent this method of learning.

While work has focused on detection methods, the literature would suggest the issue of plagiarism within education will not be solved by detection methods. Gallent claims that a change in culture would help, and other educational researchers agree [8, 28, 51, 69]. The BBC reported in 2014 that in some situations students are claiming that they have the right to cheat [98], illustrating the size of these issues are now worthy of national news. Culwin and Lancaster (2001) felt that eliminating the culture of cheating would be the best solution, commenting on the frustrations other students feel if they are aware of peers getting away with plagiarism [51]. In 2014 Ariely critiqued the current educational trends and suggested that honour codes could be an effective way to encourage students to think about honesty and their morals. Alongside this, the BBC has reported several times in recent years about the problems plagiarism pose, proving this issue has not been solved and that within the last decade there has been little improvement [98, 17, 18]. It has been suggested that the increased popularity of technology, the Internet, and the availability of methods to share work is responsible for the rise of plagiarism [50, 57, 68, 115], but the research presented over the past 15 years and evidence of students going to extreme lengths, such as passing exam answers through windows [17], show that technological advancement cannot be solely blamed for plagiarism issues.

It could be argued then, that there is little to be gained from developing tools to aid educators to detect plagiarism in non-text based work, when educating students about making moral decisions is more effective at stopping plagiarism. However when comparing the current available methods for text and non-text work, text based submissions clearly have a more established protocol for dealing with these issues. Many universities allow their students access to turnitin, so they can run their own written work through it to assess if they have inadvertently cheated and to make sure all their references are correct, however there is no way for this to happen in non-text assessments.

Despite the number of techniques presented and articles published, academic integrity is still threatened by plagiarism and there are few effective deterrent or detection techniques available to educators outside of text based subjects. This deficiency, combined with the ready availability of large 3D CAD collection in educational situations, provides an interesting research opportunity where the advantages and disadvantages of network theory and CAD can be assessed. Therefore this thesis, in chapter 4, will begin looking at the possibilities network theory can provide for analysing large collections of educational CAD parts.

2.6 Summary

The research described in this chapter is varied, taken from different disciplines and branches of science. This review has summarised the key findings and developments in the presented areas, seeking to show the links and discrepancies between them and allowing several clear observations to made from this overview.

The first is that networks, their associated techniques, tools and metrics have been proved to be useful in analysing and investigating many areas of social and physical science. Their proven functionality and robust results are key to motivating the work presented in future chapters. A brief history of graph theory, which underpins network theory, was presented in section 2.2 and the growth of network theory has been discussed in section 2.3, where the key developments were assessed chronologically. These key segments of network theory provide channels through which to investigate the use of network theory as this work progresses.

A second clear observation is that CAD has intrinsic value, making it a valid and important subject to explore and research, and this thesis will do so using network theory. The high worth of CAD data in industry is reported and can be plainly observed as CAD effects the whole of a products manufacturing life cycle. This industry is profitable and influential in society, adding to the value of this data.

It has also been shown that similarities between CAD models can be observed, particularly using shape similarity methods, which are well-researched and developed. These similarities can be seen to link CAD models, but there has been little use made of this observation. This work will make attempt to exploit and explore this, and other factors linking CAD models, using network theory.

A final consideration is that there are differing situations where large collections of CAD data is stored. CAD is observed in sizeable stores in industry and also in educational institutions. Working with ShapeSpace and Mill [14, 47, 131, 132] this investigation will be conducted in an environment where stores of data in both these areas are uniquely available for research. There is an obvious value to conducting this research with industrial CAD models, with many attempts to improve upon issues in the field. However, educational CAD models also have significant worth. The high value of CAD data is extended to CAD in educational institutions, where engineering students are trained to design and use specialist software and the issues within the field of education are also well-researched. This work will seek to assess the use of network theory related to both CAD in industry and educational situations.

Linking these well-researched and developed fields together has not been attempted before and as such this work presents novel methods for using network theory in relation to CAD data in different fields.

Chapter 3

Methodology

Throughout this work differing methods are used and presented in the relevant chapters. This chapter will briefly discuss the basics of network theory, which will be expanded upon in chapter 4, and present the software used throughout this research. This research makes use of open sourced software packages NodeXL, an add-on to Microsoft Excel, and Pajek, a free program for large network analysis. These two analysis packages, alongside currently existing ShapeSpace software, will be the main tools used throughout this investigation. During the course of this investigation ShapeSpace's software was combined with Actify's CENTRO software and as such CENTRO now include ShapeSpace capabilities. The Edinburgh Benchmark will also be introduced, a collection of around 250 CAD models created by Mill.

3.1 Graph and Network Theory Basics

Following on from the background discussion presented in sections 2.2 and 2.3 a brief overview of the basics of graph and network theory will be given to allow clear understanding of the work undertaken.

3.1.1 Nodes and edges

Networks and graphs aptly display data structure. They are composed of nodes and edges, a very simple example of which is shown in figure 3.1.

Network theory is frequently used in social science with nodes and edges widely understood in social context. In previous analysis a node (vertex, agent or entity item)

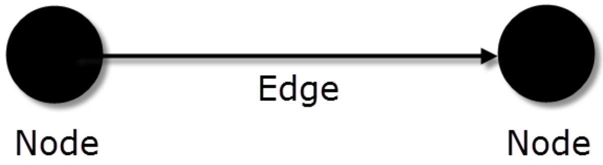


Figure 3.1: A basic directed network structure

could represent any discrete object, but most commonly they were used to represent people or social structures such as teams or groups, and often had properties attached. In this research a node may be used to represent a singe 3D part file, a sub-assembly or an assembly file. Attached to each node is attribute data, most commonly metadata linked to the file. This is used to change the visual properties of the diagrams presented. Nodes can also be used to represent collections of CAD files, by type or folder location, or other associated data, such as orders, purchases or designers. In these cases the attributes of the node change depending on what type of data it represents.

In social network analysis edges (links, ties, connections or relationships) are understood to represent the relationships between the people or social structures the graph contains. An edge can be directed or undirected, denoted by an arrow indicating the direction of the relationship. In this work edges are also used to represent the relationships between part and assembly files, though they can indicate different types of relationships. These different relationships will be stated clearly for each network and may be directed or undirected. Examples in this work include 'contains', where an assembly contains a part, or 'mirror' where two parts are mirror copies of each other. In further research edges will be used to show geometric similarity between CAD files, as well as the relationships between CAD files, orders, customers and purchases within a company's data structure.

3.1.2 Graph metrics

In this work metrics used will be calculated automatically using NodeXL. Table 3.1 presents the metrics discussed throughout this work with their explanations

Metric	Explanation
Degree (Node Specific)	A count of the number of unique edges that are connected to a node, noted as in-degree and out-degree in a directed network. A high degree shows many connections to other edges, while a low degree shows few connections.
Betweenness Centrality (Node Specific)	The number of times a node acts as a bridge between two other nodes, on the shortest path between them. Commonly used as a measure of the influence one person has on communication between others in a social network.
Closeness Centrality (Node Specific)	A measure of the average shortest distance from each node to every other node. A low closeness centrality shows a node has a more central or important position in a network.
Eigenvector Centrality (Node Specific)	A measure of how many connections a node has and the degree of the nodes it is connected to. Can show which nodes are connected to the most influential or have the most influence.
Clustering Coefficient (Node Specific)	Measures how connected a nodes neighbours are to one another.
Maximum Geodesic Distance (Diameter)	The length of the shortest path between the two nodes that are farthest apart.
Average Geodesic Distance	The average of all geodesic distances.
Graph Density	How interconnected the nodes are.
Edge Weight (Edge specific)	Value given to an edge to describe a characteristic of the link numerically.

Table 3.1: Graph metric explanations

3.1.3 Network images

When a network is constructed it is possible to also create a diagram by which to visualise the structure of the data. In this work images were produced using NodeXL and Pajek, which contain automatic layout options. Nodes can be different shapes, colours, sizes and opacities while edges can be different widths and opacities. Both can have positioned labels, but nodes can also be only labels or images if desired. Layouts can be grouped into the following categories:

- Force-directed
- Circular
- Arc
- Grid
- Tree

and this work made use of those automatically available and occasionally laid out a diagram by hand. These layout options will be discussed briefly in section 4.9 and more fully in 6.5, with relevant diagrams presented as examples.

3.2 NodeXL

NodeXL is an open source software add-on to Microsoft Excel which allows analysis of networks [176]. At its simplest NodeXL takes a list of one-to-one relationships (a network edge list) and turns it into a network. From there the software includes functions for calculating metrics, rationalising data sets and producing network diagrams.

The basic template, which is compatible with Excel 2007, 2010 and 2013, was released in 2014 and can be downloaded for free at nodeXLcodeplex.com [176]. The website hosts the software and includes documentation on use. Most commonly it is utilised for the analysis of social media networks and the website provides examples of impressive network graphs by users, along with a forum and issue log. The creators of NodeXL also published a book "Analysing social media networks with NodeXL: Insights from a connected world" which was used extensively throughout this research [80].

NodeXL is relatively straightforward to use and once an edge list is entered into the template it automatically produces a network diagram, as well as calculating metrics.

These metrics are mostly conventional and can be applied to the network diagrams as visual properties. NodeXL calculates PageRank and Modularity in addition to the metrics presented in section 3.1, but these measures were not used in this research. Notably NodeXL uses the inverse of closeness centrality in its metric report. Traditionally a low closeness centrality illustrates a node has a central or important position within a network, but NodeXL's technique means a high closeness centrality denotes the same position. This is arguably more intuitive though not conventional, as a higher value it often given to a more important position.

Gephi was also considered for use during this research [15], an alternative open source graph and network analysis software, however NodeXL was deemed more robust, and chosen to provide consistent analysis across this investigation.

3.3 Pajek

Pajek is another open source program, allowing analysis of large networks. Inputting data into Pajek can take different forms; a list of neighbours, a pair of lines or a matrix, making it suitable for different types of data analysis. Primarily built as a social network analysis tool, its creators provide documentation online at http://mrvar.fdv.uni-lj.si/pajek/ and have also written a textbook called "Exploratory Social Network Analysis with Pajek" which is in its second edition [16, 54]. In its introduction the book recommends learning by doing, expecting networks to be sparse and not dense, and aims to instruct the user in concepts and applications of network. The textbook focuses on social network analysis while instructing the user about Pajek, like a manual and provides data sets for tuition [106]. Pajek also has an extensive online community allowing users to dialogue and ask questions.

Pajek was used to produce some of the network diagrams presented in this work, however the strong focus on social sciences made the related literature difficult to use. Therefore, NodeXL was favoured throughout this thesis.

3.4 TinkerPop

During this work TinkerPop, and associated software and programming languages, were also used to create a graph database, built on an underlying network. TinkerPop is an open source graph computing framework, where data can be modelled as a network and then analysed using Gremlin, an open source graph traversal language. Figure 3.2 [153] is a network diagram created by Rodriguez illustrating the relationships between TinkerPop, Gremlin, other associated software and programming languages, and example contributors.

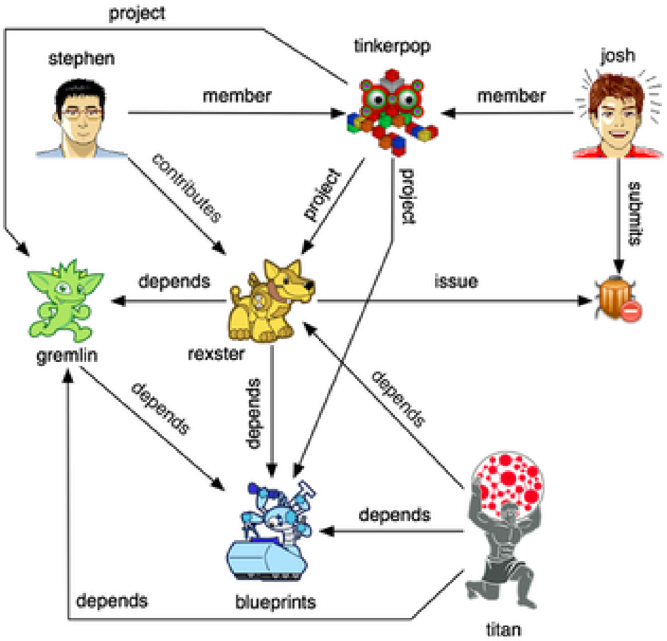


Figure 3.2: Network of relationships in TinkerPop

All capabilities provided by TinkerPop are open source and were in use at ShapeSpace while this research was conducted. Resources online include tutorials and documentation on using TinkerPop and associated technologies.

3.5 ShapeSpace Technology

The ShapeSpace software has been developed primarily by Sherlock[14, 47, 131, 132] and is used as an integral part of the data intelligence service they offer to the engineering industry. Originally focused on providing proficient shape search capabilities to large engineering and industrial design companies, ShapeSpace now prioritise equipping companies with Product Data Intelligence; new insight on their existing data with a full analysis service, to allow them to make worthwhile decisions about their products and customers, with the aim of cutting losses and improving profit margins.

The service ShapeSpace provide is based on a combination of different data analysis techniques they have developed, including their original core of highly accurate shape search software. As discussed in subsection 2.5.1 there have been many different ways found to provide accurate shape search, but ShapeSpace are among the first to effectively utilise this in engineering industry. The technology they have developed allows them to quickly and accurately detect duplicated CAD parts and group parts that are geometrically similar. Duplicated CAD designs can then be rationalised to reduce superfluous files, orders and stock, while categories of similar CAD parts can be simplified to create new, modular parts. The other technology they have developed allows them to perform further analysis on engineering BOM and design data to highlight other key areas where it is possible for companies to improve their profits and reduce their losses.

In this work the shape search and duplicate analysis portion of their software was used, now considered legacy tools. This program was reconfigured using python before use. It functions converting CAD files to STL files, which are indexed into a 'shape store'. Geometric similarity analysis can then be run on the collection. In this work the shape search and duplicate analysis capabilities of the software were used. The geometric analysis used 30 different measures to compare CAD models based on shape characteristics and have proven effective throughout ShapeSpace's work.

3.6 Edinburgh Benchmark

As discussed in subsection 2.5.2 benchmark collections of parts have been created to aid shape similarity assessment methods and have been effectively used to assess shape search methods. Most favoured and widely accepted are the Princeton Benchmark collection and the McGill Shape Benchmark. The Princeton Benchmark was purpose built for shape matching methods [166] and the McGill Shape Benchmark was created as an alternative, including CAD models with articulating parts, again with a focus on search and retrieval [206]. Both these benchmarks, as well as others [60, 97], were built to accommodate testing and experimentation of shape matching methods and therefore can be considered biased collections.

Instead of using one of these benchmarks, this work will make use of the Edinburgh Benchmark, constructed by Mill. This benchmark, rather than focussing on including CAD files with high retrieval possibilities, aims to model a typical manufacturing part collection. The part collection consists of 250 CAD parts, including assemblies, sub-assemblies and components of standard mechanical engineering designs, including nuts and bolts, screws and springs as well as more complex assemblies, such as ball valves and bike wheels. The assembly files include structured data about the make-up of the top-assembly, any inclusive relationships, and details of assembly level geometric features and finite element data. The component part files contain geometric information at various levels, including viewing meshes, history-based feature trees and other metadata, such as materials and manufacturing details. In contrast to other available collections, the Edinburgh Benchmark is a collection of CAD models that have not been rationalised or edited to include information that was not originally present in the files. The metadata included with the CAD models is original and as such the collection can be considered close to real world engineering models.

As this work seeks to examine the advances of network theory in relation to collections of 3D CAD models, it is deemed more reasonable to use the Edinburgh Benchmark as an alternative to the Princeton, McGill and other available benchmarks. It is considered that these well-established benchmarks may not provide a reliable model of a standard part collection within the mechanical engineering industry, as they have been built with respect to a specific function. They may contain more similar parts than a typical, real world CAD collection and this could have a significant effect on the results of this work. Using the Edinburgh Benchmark, built with expertise to model a real world scenario, is seen as a preferable alternative.

During the course of this research it is understood the Edinburgh Benchmark will be made available to other researchers on the on the web and the CAD models will be accessible in Solid Edge ST3 Academic, Parasolid, and STL formats. Using the open source software detailed above this will give others the opportunity to explore the CAD collections using network analysis.

Chapter 4

Investigating CAD networks

Investigation of network theory with regards to its potential impact on mechanical CAD

The research background for this work has been presented in chapter 2, where the history of graph and network theory are discussed in sections 2.2 and 2.3, while the importance of CAD data is introduced in section 2.5. Chapter 3 presented the software options available to this research as well as discussing some graph and network theory basics.

New work will be presented in this and the following chapters. This chapter will present the initial research into CAD networks, while chapters 5 and 6 will present further investigation into specific areas of network theory, in the context of CAD in education and CAD in industry.

This chapter will discuss some of the most common and prevalent theories of the science of networks, drawing on the background presented in section 2.3 and will explore their relevance, advantages and disadvantages for 3D CAD in mechanical engineering and consider real world interpretations. The leading theories and network developments will be discussed and examples of how these can be applied to networks of differing CAD collections will be explored. Applications are found and discussed in various settings and subsections, including CAD in industrial and educational situations. Associated metrics and measures will also be explored and conclusions presented.

4.1 Introduction

This research was inspired by the work of Mill [14, 47, 131, 132] and the books published by Barabási (2002) [12], Buchanan (2002) [37] and Watts (2003) [194] were used. A summary of the key findings in the area has been presented in section 2.3. As networks have been extensively studied in the field of social science, little is directly applicable to mechanical engineering designs, however there is a wide consensus that networks are useful tools with which to explore existing data structures.

Unsurprisingly the popular works by Barabási, Buchanan and Watts have been the foucs of many critiques. In one such article comparing the Buchanan's and Barabási's books, Aldana-Gonzalez criticises both authors for writing for a lay audience and applying a popular, reductionist approach to network science. She states that "...both authors overstate the scope of the new science of networks in an attempt to connect scientific disciplines and to unify the law of networks" [4]. However Aldana-Gonzalez's statement neglects the concept that networks, by nature, are able to link previously unconnected areas and disciplines, and to vilify Barabási's findings based on writing style seems ungenerous. Indeed "Linked" by Barabási is easily digestible and comprehensibly written, so that a new researcher to the field can quickly and easily comprehend the concepts and reach a good understanding of the scope of network theory within a relatively short time. Despite his pragmatic writing, Barabási is still considered one of the main contributing authors to the development of network theory today and as such his work deserves respect.

The analysis in this chapter will follow the sequence of key advances in network theory as noted by Barabási. It will begin with random networks in section 4.4, a concept introduced by Erdős and Rényi, and continue on to six degrees of freedom in section 4.5, discussing how Milgram's work [130] relates to mechanical engineering networks. The other areas that will be discussed will include small worlds (Watts and Strogatz [197]), hubs and connectors (Barabási [12]) and weak ties (Granovetter [77]).

As the theory of networks has most frequently been used within social science, it has not been applied to groups of 3D Computer Aided Design (CAD) models in a mechanical engineering context prior to this work. As discussed in chapter 2, within mechanical engineering there are many different places where CAD models can be found but most notably they are used in the manufacturing design industry and educational institutions. The types of files can vary hugely, from simple, straightforward components, such as nuts and bolts, through to large scale assemblies, for instance full aeroplanes. Similarly the larger collections containing these individual files will differ depending on context, scale of business and industry sector.

It was decided that taking the main themes and developments from the history of network theory and analysing how each may be applied to differing types of CAD part collections found in the real world scenarios would be an effective, pragmatic and thorough way to begin this analysis. Different CAD collections could be modelled and then compared and contrasted to explore the uses of network theory within engineering design.

4.1.1 Networks modelling CAD collections

There are two distinct areas CAD models are found in as discussed in section 2.5, namely industry and education, and part collections from these two areas may look quite different. It is likely that collections from industry would be much larger, with many more designers, while collections from educational institutes will be smaller, with fewer contributors. Each student may have a part collection that reflects that of a single industrial designer who will have a small range of parts related to their area of expertise. Every CAD collection, regardless of location or ownership, will likely consist of components, sub-assemblies and assemblies.

When considering how to model these CAD collections in network form, it is vital to consider the function of the nodes and the edges. It would seem natural to model a CAD collection where the components were represented by nodes and the links between them could be a characteristic link. However, nodes could also be used to represent sub-assemblies and assemblies or entire designs, as well as 2D drawings instead of 3D models. Edges linking the components could illustrate that parts were created by the same designer or that they were contained within the same assembly (a 'where used' or 'contains' relationship). Edges could also represent CAD models that were linked by shape similarity, function, size, author, profitability, cost or any other metadata available with a model.

In order to clearly and fully analyse these different situations five CAD collections were modelled. It was decided that some should be simple, to allow for clear and straightforward analysis, while others should be larger and more complex for comparison. These 5 networks were numbered for clarity during discussion and created from the following CAD collections:

- **Network 1**

  A simple assembly structure network.

  The CAD collection used to build this network was an assembly of a simple valve structure and all the associated components, from a student model at the University of Edinburgh. In this network a node represents a component, sub-assembly or assembly and an edge represents a 'contains' relationship as traditionally understood within CAD methods and as such this network can be termed an assembly structure network.

- **Network 2**

  A simple assembly network of two semantically similar assemblies.

  The CAD collection used to build this network contains two similar valve assemblies, which share some simple components. Once again, in this network a node represents a component, sub-assembly or assembly and an edge represents the 'where used' relationship. As the two top-assemblies share components this network will be more complex than network 1, while still being comparable in size and simplicity. These two valve assemblies have been taken from the Edinburgh Benchmark, presented in section 3.6

- **Network 3**

  A unimodal network of the Edinburgh Benchmark

  The CAD collection used to build this network is the Edinburgh Benchmark, a repository of parts created to provide industry standard, real world models. In this network a node will represent a CAD file, which may be a single component, a sub-assembly or an assembly, and an edge will represent the 'where used' relationship. This network is built from the whole Edinburgh Benchmark, so will include Network 2.

- **Network 4**

  A network of a CAD collection from education.

  The CAD collection used to build this network is made up of submissions from a CAD course at the University of Edinburgh. The parts have been anonymised, but otherwise remain unedited for this analysis. In this network a node represents a student, or more accurately a single node represents all the relevant CAD files belonging to a single student, and an edge represents the similarity between the students' files. In this case if one node is linked to another, that shows there is similarity between the work those students have submitted. It is important to note that only relevant CAD files from a single student, those deemed suspiciously similar to other files within the collection, are represented by a node, not the entirety of a student's hand-in portfolio. The similarity of the CAD files in this collection was found using ShapeSpace's legacy software, presented in section 3.5.

- **Network 5**
  A network of a large, industrial CAD collection.
  This network is built using a sizeable CAD collection of over 13,000 models, which is real data anonymised from a large industrial client of ShapeSpace, company A. The CAD files have not been edited, other than to remove the company's identity, so the models contain all original metadata. In this network a node represents a component, sub-assembly or assembly and an edge represents a 'where used' relationship. Within the company this relationship was called 'item contains item' and refers to the standard assembly structure using terms more commonly found in programming than in the mechanical engineering industry.

Each of these networks was created using the NodeXL open source software discussed in section 3.2.

4.2 Initial Mapping and Observations of Networks 1-5

Once networks 1-5 had been constructed from the CAD data, they were analysed individually and prepared for comparison. In each case the relationships between CAD files were entered into a NodeXL spreadsheet and then assessed. Visualisation was arranged to maximise visual clarity, using the forced layouts included in NodeXL or by hand.

Network 1 was built from a simple assembly structure, a turbine modelled by a student at the University of Edinburgh. This turbine assembly had five models that were combined to make the final top-assembly. Of these five models, one was a sub-assembly, which contained 4 other models.

Figure 4.1 shows the initial network diagram that was produced from this assembly structure where some metrics have been mapped to visual properties of the diagram, while figure 4.2 shows the same network with further metrics mapped to visual properties. The layout was arranged using the Sugiyama layout, which forces the network into a layered diagram where the nodes are arranged in rows. This is a suitable layout

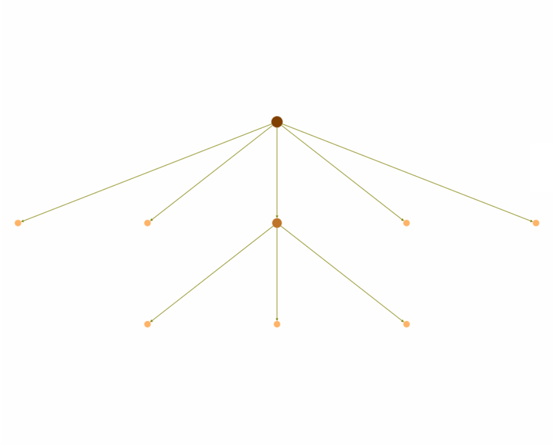


Figure 4.1: Network 1: Initial mapping

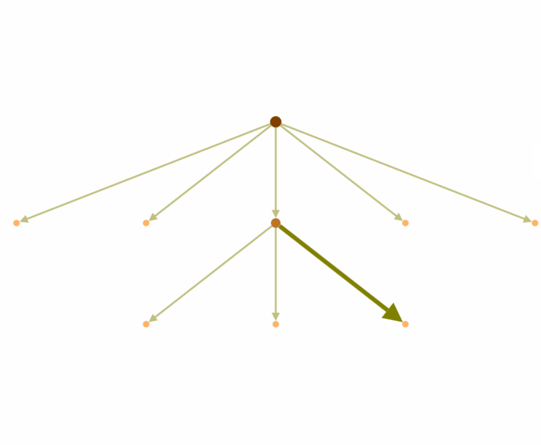


Figure 4.2: Network 1: Adjusted visualisation

for this simple assembly, as it reflects a classic tree structure, which is well understood and commonly used within CAD, and highlights the levels of CAD models the layers represent. In both figures the node size and colour is set to show the out degree of the node. This results in two larger, darker nodes, which clearly represent the top-assembly and sub-assembly within the network. This interpretation is helped by the Sugiyama layout. In figure 4.2 the edge weights have been edited to represent not only the 'contains' relationship but also how many times a part is contained within the containing CAD model. Figure 4.2 clearly shows there are many of one particular part contained within the sub-assembly in this network. Notably multiple edges cannot be displayed in NodeXL and so must be indicated using edge weight. In figure 4.1, where there are multiple edges between the two nodes is not evident. However in figure 4.2 the use of a heavy line for these edges clarifies the situation.

Network 2 was also laid out using the Sugiyama layout, however as there are more nodes (23 nodes in network 2 compared to 9 nodes in network 1) figure 4.3 shows how this layout rapidly becomes unsuitable for larger networks. The nodes are represented by images of the actual CAD models in this network diagram, instead of simple circles, to allow for easy interpretation of the diagram. In this network many of the sub-assemblies and single parts are common to both top valve assemblies, seen by the multiple edges traversing the network.

While the components can be clearly seen, the larger number of edges and vertices makes this layout unclear. As there are only three layers to the assembly structure, the diagram becomes wide and squat, compact and less intuitive. To improve this layout the

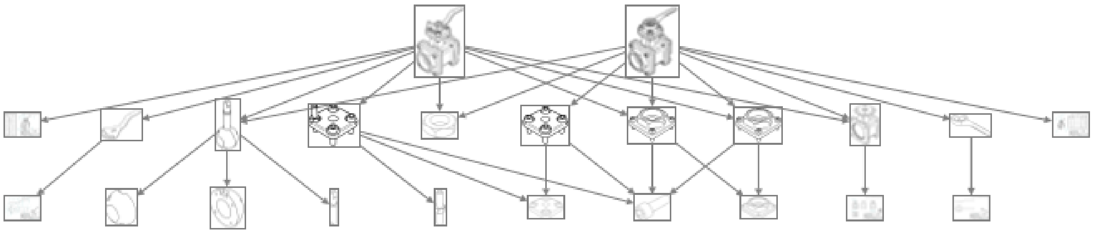


Figure 4.3: Network 2: Valve assembly in Sugiyama layout

network was arranged by hand, by Mill, as shown in figure 4.4. While this version of the network diagram does not have the strict layers of the Sugiyama layout, it maintains the structure resembling assembly levels, similar to those found in a traditional tree diagram.

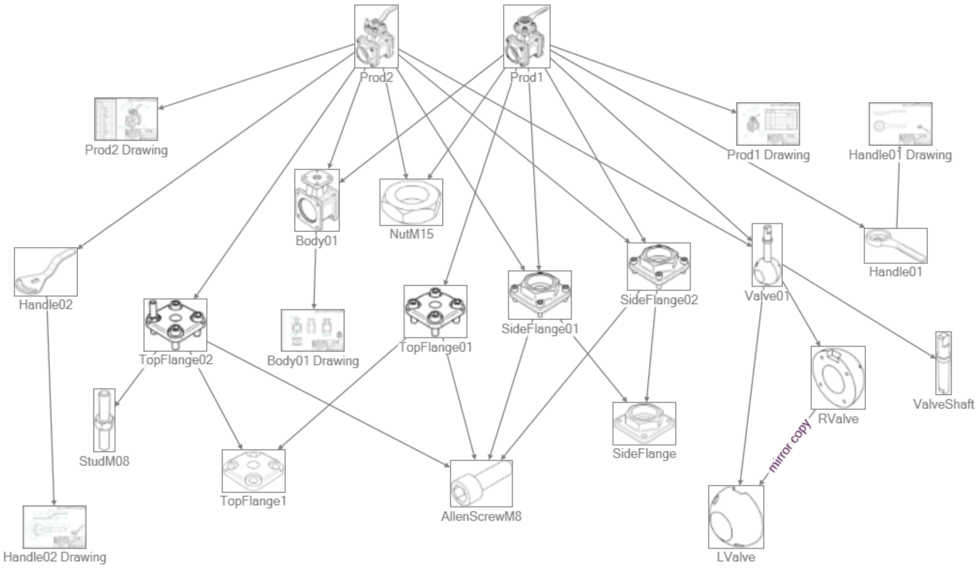


Figure 4.4: Network 2: Valve assembly arranged by hand

Figure 4.4 shows the two top valve assemblies connected to technical drawings of the models, as well as all the CAD components and sub-assemblies that they contain. This is true for other parts in the network too. There is also an edge displaying a different relationship between two CAD parts; the right valve and left valve are mirror copies of each other, and this information has been included in the network. These different links make this network more complex, modelling a real world engineering design situation well, while the network as a whole remains relatively small and simple. These added nodes and edges make the network multimodal, rather than unimodal, and this may affect the associated metrics and other measures when comparing network 2 to others.

Network 3 is built from the Edinburgh Benchmark, and so is much larger than networks 1 and 2. The 250 CAD files are represented by the nodes and figure 4.5 shows an initial mapping of the network created with edges representing assembly structure.

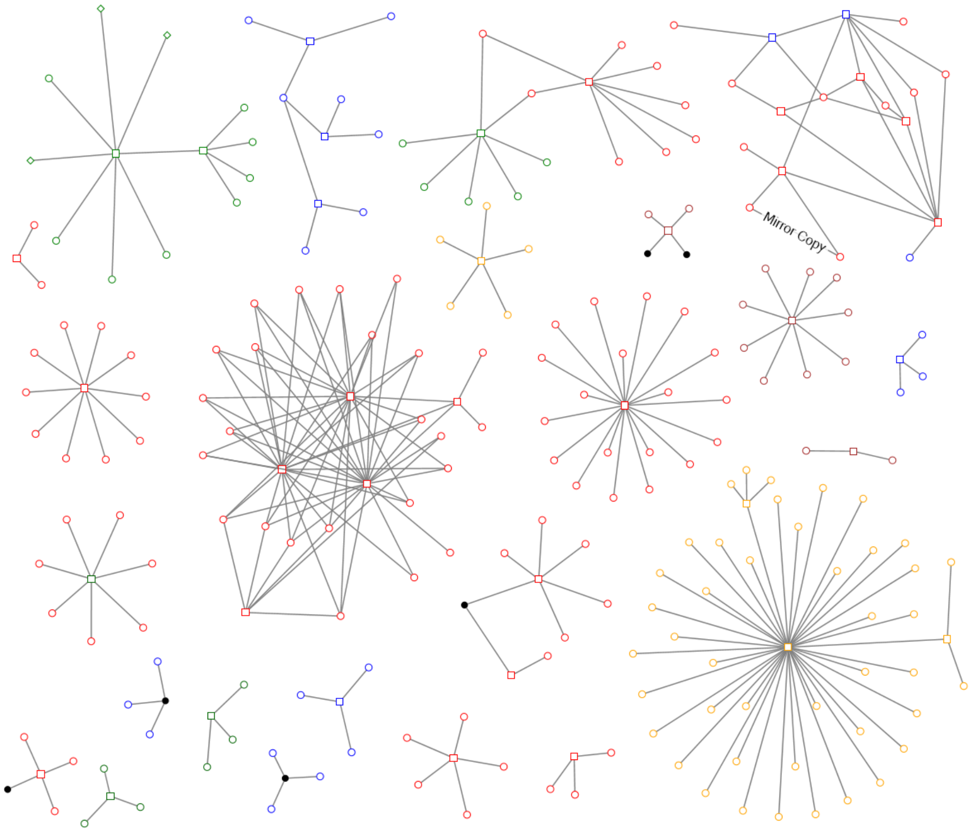


Figure 4.5: Network 3.1: Edinburgh Benchmark parts modelled with assembly relationships. N.B. This is referred to as network 3.1 as later in this chapter different visualisation of the data will be presented, called 3.2 and 3.3

The graph in figure 4.5 shows that the network is largely made up of individual clusters, each representing one or more related assemblies. Clearly visible in the top right-hand corner are the two valve assemblies that were used to create network 2. Square nodes identify assemblies and circular nodes identify component files, while subassemblies are also quadrilateral shapes. The colours were chosen to represent the status of the associate CAD file within the collection, where green shows a part is ‘available’, yellow ‘in review’, blue ‘released’ and red ‘baseline’. Black nodes show there is no CAD file associate with the Bill of Material’s (BOM) entry for that part.

The network shown in figure 4.5 was laid out by hand to clearly present the differing clusters and, while displaying information about the assemblies within the Edinburgh Benchmark plainly, this particular network construction may be of little use to this investigation. It is unlikely that any global metrics, such as average geodesic distance, could be meaningfully calculated however this network structure may be useful in characterising a large collection of engineering design models.

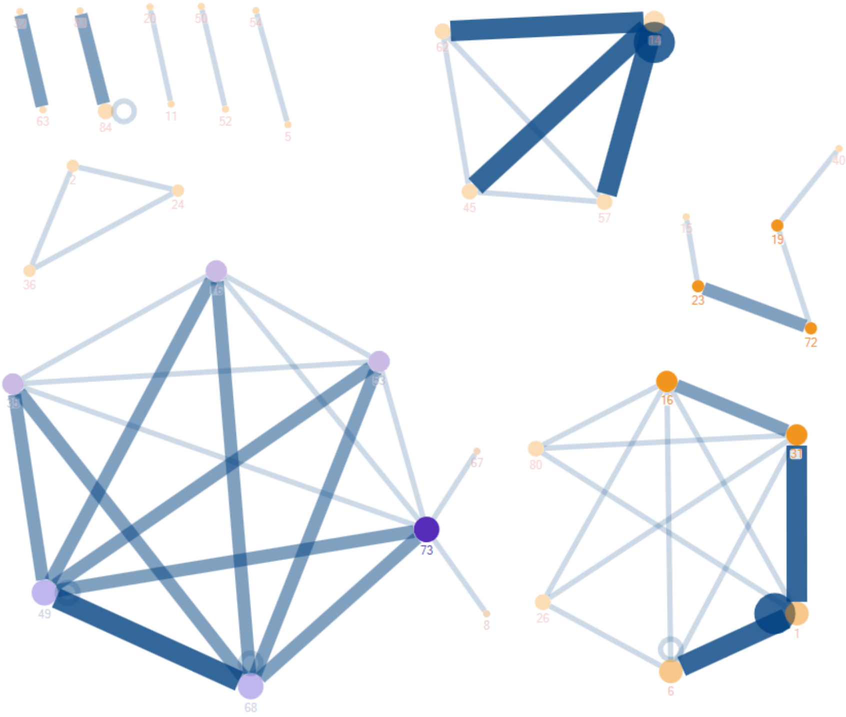


Figure 4.6: Network 4: Students' CAD data linked by shape similarity.

Network 4 shown in figure 4.6 was built from a larger CAD collection, drawn from an educational setting. It is important to note that an individual node does not represent a single CAD file, rather a 'folder' of CAD files belonging to a single student and in contrast to networks 1-3.1, the edges linking nodes represent similar files contained in one 'folder'. Despite the source CAD collection having more files that the Edinburgh Benchmark, the resultant network contains only 36 vertices, due to the data preprocessing before being used to construct network 4. This network was laid out by hand as no suitable forced layout could be found. The metrics calculated were mapped to visual properties of the graph, to investigate which of them might most helpfully reveal interesting characteristics and properties. Again the network contains several separate clusters, some of which are highly connected. This network requires careful consideration and is not directly comparable to the diagrams of network 1-3.1 (figures 4.2, 4.4 and 4.5), as the nodes do not represent single CAD files and the edges show a different relationship.

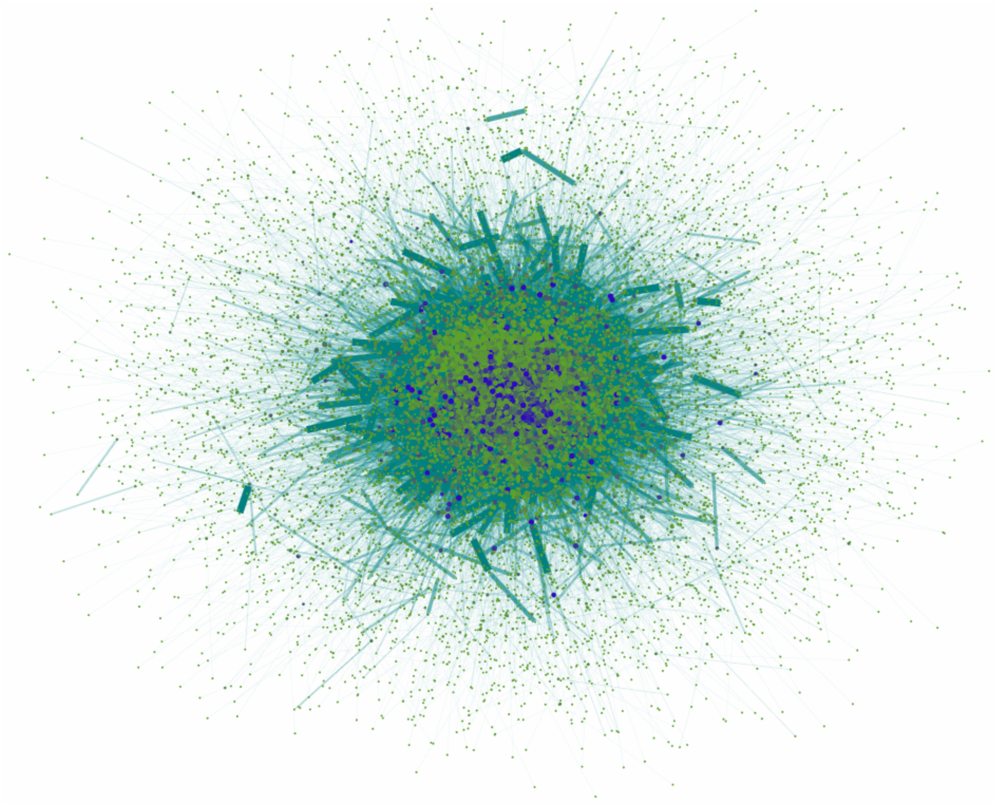


Figure 4.7: Network 5: Large industrial CAD collection

Network 5 was built with a CAD collection of over 10,000 files from Company A. The data was a whole database from a company's PLM system, made available to ShapeSpace when they were performing Data Intelligence analysis for the company. The parts are unedited and the relationship modelled was the 'item contains item' structure, which is semantically equal to the 'contains' relationship modelled in networks 1, 2 and 3.1. The network, shown in figure 4.7, was arranged using the forced layout Fruchterman-Reingold, as the large number of nodes make it impractical to arrange this graph by hand. While being visually striking, it is interesting to note the differences between this very large collection and network 3.1, showing the assembly structure of the Edinburgh Benchmark. It is very likely that this network contains some isolated assembly structures, however many of the singular CAD components are used multiple times in the company's assemblies, so this large collection may be more connected than network 3.1 and this should be discernible from the metrics of both networks.

There are several striking differences between the five networks visualised here. Networks 1, 2, 3.1 and 5 are created using 'contains' relationships, a typical relationship found within CAD data, while network 4 is created using shape similarity methods to generate the relationship modelled. Despite this, the larger collections modelled in networks 3.1 and 4 have clearly isolated clusters within the network, as network 5 presumably also does, though these cannot be identified visually. Networks 1, 2, and 5 also are directed graphs, where the edges travel specifically from one node to another, while networks 3.1 and 4 are undirected graphs. Network 3.1 could be a directed graph if the assembly structure modelled was converted into 'contains' relationships, but network 4 could not be transformed into a directed network. The similarity between files found in the shape analysis is an equal relationship and could not accurately be represented by a directional edge. Notably Network 4 is the only network to contain self-loops, where a node is linked to itself. In this case the edge is showing there is similarity within a student's 'folder' of submitted work.

Another notable difference between the networks is their mode, an attribute used to categorise networks. A mode denotes how many different types of node a network contains. A network is said to be unimodal if it contains one type of node, a bimodal network is one where there are two types of nodes and a multimodal network has several types of nodes and edges. While networks 1, 3.1, 4 and 5 have only one type of edge, networks 1, 3.1 and 5 are arguably bimodal, network 2 is arguable multimodal and network 4 is unimodal. It could be argued that networks 1, 3.1 and 5 are unimodal, as their nodes represent CAD data. There has been discussion suggesting that global metrics do not hold true for bimodal or multimodal networks as they, by definition, are measures that can only be applied to one type of node. As such, for the purposes of this work the CAD data used to build them may allow a unimodal view to be held, particularly when assessing associated metrics from a network analysis viewpoint, however it is notable that the CAD data modelled by nodes within these networks is not uniform, as a single CAD part and an assembly are completely different from an engineering viewpoint. For this reason, these networks will be viewed as unimodal for some of the subsequent analysis, while arguably untrue from an engineering perspective.

Another notable network with regards to mode is network 2, where nodes represent 3D components, sub-assemblies and assemblies, as well as 2D drawings, and the edges linking them are 'contains', 'mirror copy' and 'drawing of' relationships. In this case, while the 2D technical drawings are not 3D components, they can still be regarded as generic CAD files. From an engineering viewpoint it makes logical and technical sense to have 2D drawings linked to their corresponding 3D file, however from a network analysis standpoint, it may be preferable to consider these drawings simply as CAD files when examining this network. Including these files may provide different possibilities, particularly when considering engineering search, however they may impact the global metric measurements of the network, so must be carefully considered.

Each of the created networks 1-5 is unique and, outwith Mill's research group, no similar network analysis has been performed on CAD models; as such the networks presented here are entirely novel. Networks 1-5 will now be compared and analysed according the network theory developments, layouts and metrics.

4.3 Metrics and Measurements

In order to clearly compare the metrics of the networks 1 to 5, table 4.1 shows the basic associated metrics for each. The size of each network can now plainly be seen, shown not only by the number of nodes and edges each network has, but also the global measures. Here the mathematics of individual metrics will not be discussed, as these are well established and commonly used within the field. For further discussion of metrics refer to section 3.1.

Network 5 is obviously the largest network by a sizeable margin. It has the largest number of nodes and edges, as well as the largest maximum and average geodesic distances. Network 5 also has a very low graph density. Graph density is a measure of how many edges a graph contains compared to how many it would if it was fully connected. It can be seen that while network 3.1 and 4 are sparse, they are comparably dense to network 2. Network 5 is the least dense graph and network 1 is the most dense in this comparison, while still having a relatively low density measure. Density could be an indication of how easy it would be to search for a CAD file within a collection, or it could be a characteristic measurement of large CAD collections. If a network of an industrial CAD collection was dense it could indicate that parts were well used, while a spare density measure could indicate low part use.

Also instantly notable are the two networks, 3.1 and 4, which have a number of separate components. This was plainly observable in the network diagrams, shown in figures 4.5 and 4.6, but is interesting to note the effect this has on graph metrics. These separate components invalidate the measures of maximum and average geodesic

Metric	Network 1	Network 2	Network 3.1	Network 4	Network 5
Type of graph	Directed	Directed	Undirected	Undirected	Directed
Number of nodes	9	23	245	36	13125
Number of edges	8	32	275	53	65007
Number of self-loops	0	0	0	6	0
Maximum geodesic distance (Diameter)	3	6	5	4	13
Average geodesic distance	1.87	2.62	2.00	1.07	5.01
Graph density	0.11	0.06	0.01	0.07	0.00038
Separate components	1	1	24	10	1

Table 4.1: Basic metrics of each network

distance, as it is impossible to travel from some nodes from a given starting node in these network structures. As such these measurements reported in the table relate to one of the clusters, not to the graph as a whole. Despite this, both diameter measures refer the diameter of the largest component in the network, which may be of use.

Network 4 is the only network to have self-loops, and this is because of the data modelled. As the edges represent similarity between the stores of students' work, it is conceivable that a student may have submitted multiple files that are geometrically similar to each other. It is highly unlikely that a self-loop would be found within a network modelling assembly structure, as a CAD file could not 'contain' or be linked to itself. If this kind of edge was present in networks 1, 2, 3.1 or 5 it would indicate a clear error in the data, and as such could be very useful in assessing large part collections linked by assembly data.

In order to further assess metrics and how they can be applied to networks of CAD collections, table 4.2 details the common interpretation of network metrics and presents a suggested interpretation of these metrics when measuring a network, such as network 1, 2, 3.1 or 5, of a CAD collection modelled by assembly structure. These metrics are well understood within social science, but have never been used or applied to mechanical engineering design situations or networks modelling CAD collections before, as such the recommended definitions presented in table 4.2 are entirely novel.

Table 4.2 presents new definitions for network metrics, particular to network 1, 2, 3.1 and 5. The node specific measurements are not visible in the network diagrams, unless mapped to visual properties. This was effectively done for network 1, shown in figure 4.2 where the nodes size is determined by the degree of the node. While this is clear

Metric	Classic Interpretation	CAD Assembly Networks
Degree (Node Specific)	A count of the number of unique edges that are connected to a node, noted as in-degree and out-degree in a directed network. A high degree shows many connections to other edges, while a low degree shows few connections.	How many CAD models are connected to this one. In-degree shows how many assemblies a part is used in, while out-degree notes how many different parts an assembly contains.
Betweenness Centrality (Node Specific)	The number of times a node acts as a bridge between two other nodes, on the shortest path between them. Commonly used as a measure of the influence one person has on communication between others in a social network.	How many time a CAD part bridges two others. May provide a measure of how important a part is in the network.
Closeness Centrality (Node Specific)	A measure of the average shortest distance from each node to every other node. A low closeness centrality shows a node has a more central or important position in a network.	A measure of the position of a part within the collection.
Eigenvector Centrality (Node Specific)	A measure of how many connections a node has and the degree of the nodes it is connected to. Can show which nodes are connected to the most influential or have the most influence.	A measure of the most influential parts.
Clustering Coefficient (Node Specific)	Measures how connected a nodes neighbours are to one another.	A measure of how connected assemblies and sub-assemblies containing a part are to each other.
Maximum Geodesic Distance (Diameter)	The length of the shortest path between the two nodes that are farthest apart.	The number of steps between the two CAD models that are the most unrelated or the maximum span of the collection.
Average Geodesic Distance	The average of all geodesic distances.	The size of the collection of parts.
Graph Density	How interconnected the nodes are.	A measure of how connected the CAD collection is. Could indicate how well parts are used.
Edge Weight (Edge specific)	Value given to an edge to describe a characteristic of the link numerically.	Determined by the characteristic of the link, here how often a part is used within an assembly.

Table 4.2: Classic interpretation of graph metrics and novel interpretation with respect to a CAD collection network

in such a small network, when applied to a larger network, such as network 5 in figure 4.7, the effect is reduced though some key nodes can still be identified. Representing these node specific metrics with visual properties is an effective aid to communication however if too many are applied, the diagram can become unclear and the meaning and interpretation become confused. This is illustrated by network 4, where several node specific metrics have been mapped to visual properties, producing an image that needs a key to be correctly interpreted (see figure 4.6).

The global measures of maximum and average geodesic distance and graph density are suggested as metrics to indicate key characteristics about CAD collections. The diameter of a network may provide a new indicator of the size of a CAD collection, while the density of a network may clarify how connected a data set is. The size and how connected a collection is may have implications for search functions, where the average geodesic distance might suggest an average number of steps between sought files.

Edge weight was assigned to a measure in network 1 and 5, where it indicated the number of times a part is contained within an assembly one level up from the component. The edge weight in network 1, shown in figure 4.2, is not mapped in a representative ratio, and as such could give a false impression of the number of parts contained within the next level assembly, whereas in figure 4.7 the larger number of parts allows for an even distribution of edge weight, and the image shows an interesting distribution of part use, which can be compared to other global metrics.

It can also be seen that these metrics may be instrumental in characterising collections of student CAD data and further work, presented in chapter 5 will continue to assess this.

4.4 Random Networks

Erdős and Rényi were mathematicians who first discussed random networks in 1959. They introduced a model for creating random graphs based on probability and used it to discover properties of typical graphs. In this work so far no random graphs have been presented, as networks 1-5 model specific data structures. In order to discuss random networks, however, Mill used Pajek to create two random networks from the Edinburgh Benchmark collection, one dense and the other sparse. Pajek randomly assigned links between nodes, according to a probability distribution and the two resultant networks are shown in figures 4.8 and 4.9.

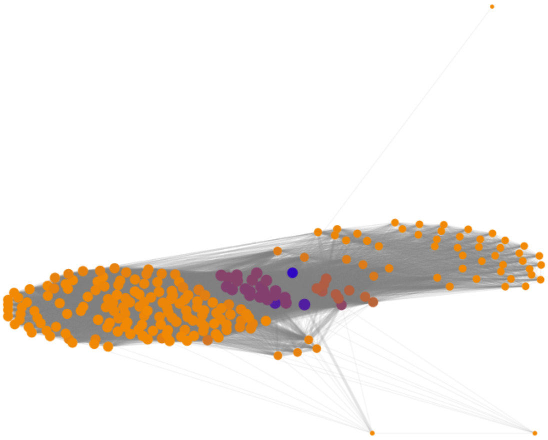


Figure 4.8: Network 3.2

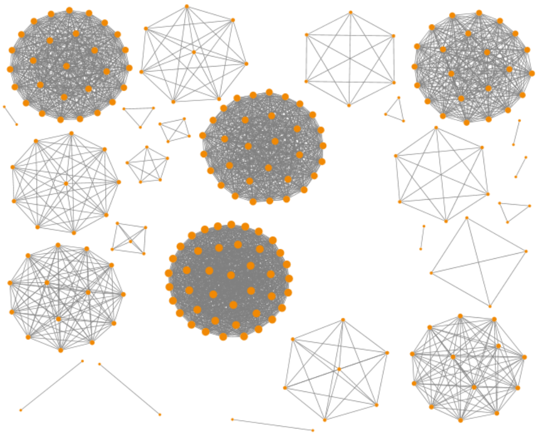


Figure 4.9: Network 3.3

Both graphs have had the same metrics mapped to visual properties to allow comparison. In both diagrams edge width and opacity is determined by edge weight, while degree was mapped to node size and betweenness centrality was mapped to node colour, with orange showing a low measure and blue representing a higher betweenness centrality. Figure 4.8 shows network 3.2 in a Fruchterman-Reingold layout while figure 4.9 shows network 3.3 arranged by hand. Immediately obvious is the structural difference between the dense and sparse networks. Network 3.2 appears highly connected, while network 3.3 appears as separate components. Also evident is the difference between some key metrics. While both networks contain different sized nodes, indicating there are differing values of degree between nodes, network 3.3 is shown to have no variation in edge weight, as all edges are the same width and opacity, or in betweenness centrality, as all nodes are the same. This is a function of the sparse network generated.

The only apparent similarities between network 3.2 and 3.3 is that they model the same data on their nodes and are both randomly generated, undirected graphs. The significant variation in their metrics is presented in table 4.3

The randomly generated graphs have resulted in network 3.2 containing one more node that network 3.3 and a vast difference in the number of edges, with network 3.2 having almost 8 times more edges than network 3.3. This is reflected in the graph density measures, where network 3.2 is shown to be almost 8 times denser than network 3.3. These are reliable measures, due to the differing probability distributions used to generate the networks. Other key difference are that network 3.2 contains a small number of self-loops, showing that randomly generated networks can contain self-loops.

Metric	Network 3.2	Network 3.3
Type of graph	Undirected	Undirected
Number of nodes	231	230
Number of edges	18338	2329
Number of self-loops	2	0
Maximum geodesic distance (Diameter)	3	1
Average geodesic distance	1.31	0.95
Graph density	0.690	0.088
Separate components	1	25

Table 4.3: Metrics of networks 3.2 and 3.3

Also table 4.3 shows that network 3.3 has 25 separate components and as such the maximum and average geodesic distances are not accurate global measures.

An effective illustration of the difference in density and connectivity between the two networks is shown in figures 4.10 and 4.11 where networks 3.2 and 3.3 are shown in a forced circular layout.

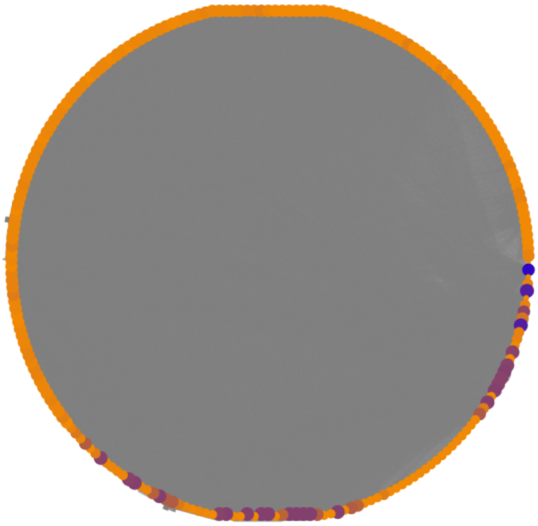


Figure 4.10: Network 3.2 in forced circle layout

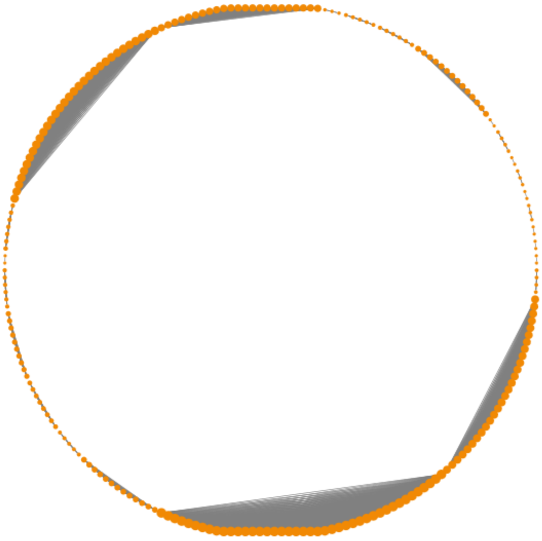


Figure 4.11: Network 3.3 in forced circle layout

From figure 4.10 it can be seen that network 3.2 is so highly connected the edges become impossible to distinguish from each other, while figure 4.11 shows how sparsely connected network 3.3 is, with few highly connected segments being separate from the rest of the nodes.

While networks 3.2 and 3.3 provide opportunities to further investigate general networks, they do not have a significant meaning when exploring the uses of network theory in relation to CAD data. These randomly generated network structures do not reflect any meaningful data with regards to CAD or mechanical engineering and it is not possible to draw conclusive interpretations from the metrics generated. As such this work will not consider random networks, or networks 3.2 and 3.3, any further.

4.5 Six Degrees of Separation

Milgram wrote of a social experiment in the 60s, in which he attempted to measure a line of acquaintance linking any two randomly chosen persons. From his experimentation, Milgram concluded that while at the time many social studies suggested individuals were isolated from the rest of society, this study showed all persons were bound together in a tightly knit social fabric [130]. He seemed to conclude, amidst surprise from his peers, that an average of 5 intermediaries are needed to link any two randomly chosen individuals. Following Milgram's paper, the phrase 'six degrees of separation' has become commonplace and well used. Much research has been conducted, taking this measure as correct, and there has even been a play written by John Guare, later made into a film in which Will Smith plays a disconnected youth who changes the lives of elite couples in New York [163]

In "Linked" Barabási questions Milgram's measure and writes of experiments performed to determine whether it holds true in other networks. Continuing the line of investigation, Barabási set out to determine the degrees of separation on the Web. In 1998 the Web as modelled by a connected network was estimated to have around 800 million nodes and the experiments concluded the diameter of the Web was 18.59 [3], otherwise expressed as 19 degrees of separation. Barabási goes on to report this has been determined as true in other areas, stating that the molecules in cells are separated by only 3 chemical reactions and authors from differing fields of science are 4 to 6 collaborations apart, showing many systems modelled by networks are more closely linked that previously thought. In fact the Web with 19 degrees of separation is reported as the largest [12]. These differing measures seem to suggest that relatively small sizes of separation should be found in networks representing any connected area or system. It is worth noting that while 19 seems a much larger measure than 6, when the size of the investigated system and resulting is taken into account, this difference appears negligible.

It is of interest to assess if, within a CAD collection, a single CAD part could be reached within 6 steps from another. Assessing networks 1-5 it is obvious that there are several of the networks that will not conform to this measurement, as they are not connected enough to effectively allow a diameter, also called maximum geodesic distance, measure. Using NodeXL the metrics were calculated for networks 1-5 and the following diameters were found:

- Network 1: 3
- Network 2: 6
- Network 3.1: 5 (see explanation below)

  Notably there are 24 connected components (individual clusters) in the network. The largest number of edges in any cluster is 71, while the largest number of nodes is 46 and these are not the same clusters. The diameter calculated for this network is the largest for any one of the single clusters, not for the network as a whole. It is not possible to calculate the diameter due to the disconnected clusters present, and as such the true diameter of the network must be reported as infinite.
- Network 4: 4 (see explanation below)

  Again, as with network 3.1, this network has several individual clusters of smaller connected groups. In this case there are 10 connected components, the largest measured as having 8 nodes, and the largest edge number is 19. In this case these metrics correspond to the same cluster.
- Network 5: 13

  Interestingly this network is one, large connected component, not many smaller ones as originally thought. There are no individual clusters, instead all the CAD files are linked, the furthest being 13 away from each other. This is notable as it shows company CAD data may be more linked than previously thought. This could be due to a company producing many of the same kind of product, so many top-assemblies use similar components. This would be especially true of small components such as screws and bolts, which may be common to most assemblies. In this way, the network of parts would be well-connected, via these small, vital components.

As networks 3.1 and 4 are not connected networks, but collections of separate clusters, the measure of size of the network, using diameter is arguably useless. Comparing networks 3.1 and 4 is also of very dubious value, as they represent significantly dissimilar CAD collections. The structure of the two networks means that other metrics are more suitable to apply to them, and as such they will be excluded from the rest of this analysis.

Network 2 has a diameter twice that of network 1 and this is expected as it is much larger. A simple network, such as network 1, with very few nodes would be expected to have a small diameter while network 2 would be expected to have a larger diameter, simply due to the larger number of nodes and edges. However the diameter of network 1 is arguably large, given its small size. This is particularly evident when compared to network 5, which has a diameter of 13. While this diameter is much larger, over four times larger, network 5 is not only four times but much larger than network 1. Comparing these three networks, as they all model assembly structure data, it can be seen that on a small, individual assembly scale the diameter of a network is relatively large, while for a sizeable collection the diameter is relatively small.

This does not agree with Milgram's theory, instead agreeing more with Barabási's findings. The context of these networks is the key difference between previous measures and those recorded here and so it can be concluded that networks of CAD files do not agree with the 'six degrees of separation' theory.

For CAD data, the main application of this measure would be relevant to search capabilities. If a network of CAD files had a small diameter, it could be seen that searching for one CAD file by starting with one and exploring those close to it, the network would be straightforward to search. Networks 1 and 2 are small enough that a designer wouldn't need assistance in searching through that data, but to search within network 5 a designer would require help. If a company's data was similar to that modelled in network 5, it may be possible for the data from a PLM system to be used to create a searchable network. However if their data is structured as that in network 2, it may not be a reliable method.

A key problem with this line of investigation is that it has been assumed from Milgram's work that he proved the 'six degrees of separation' theory, however his experiment cannot be factually concluded this way. While it is now commonly reported that everyone is only 6 degrees of separation away from anyone else in the world, this is not true and Milgram's conclusions are based upon some questionable assumptions. Milgram's experiment had a low accuracy, as there were many letters that did not reach his target person. Of the 42 letters that reached the stockbroker the median number of steps was 5, however for the other 118 it can be said that the steps between the start person and the target person were infinite, as they never arrived. While Barabási used Milgram's findings to inspire further work, there has been little work done to confirm these findings. Studies reported in this section discuss the ongoing experimentation. While Milgram's finding are relevant only in social science, it is interesting to note that the diameter of a network may not have as much significance as previously thought. In this case the maximum geodesic distance does provide a novel measure of the size of a CAD collection and in this way is significant for this work.

4.6 Small World Theory

According to Watts and Strogatz all networks fit neatly into one of three categories; Ordered, Small world, or Random [197]. They claimed that there are no naturally occurring ordered networks and, in nature, no truly random networks exist either, instead existing only as computer generated artefacts. The findings in this chapter corroborate this observation as the randomly generated networks, networks 3.2 and 3.3, presented in section 4.4 are computer generated and bare little resemblance to networks 1, 2, 3.1, 4 and 5, which model real CAD data structures.

Most networks have been shown to be of the small world variety, discussed in subsection 2.3.3, where clustered nodes are well-connected to each other, with longer, more randomly occurring links connecting them to the surrounding clusters. To measure this, Watts and Strogatz introduced using the clustering coefficient as an effective measure of this. Table 4.4 displays the minimum, maximum and average clustering coefficient for nodes within networks 1-5.

Clustering coefficient	Network 1	Network 2	Network 3.1	Network 4	Network 5
Minimum	0	0	0	0	0
Maximum	0	0.5	1.0	1.0	1.0
Average	0	0.046	0.009	0.487	0.009

Table 4.4: Clustering coefficient metrics for each network

Network 1 has no clustering coefficients to compare, as all measures are calculated as 0. This may be indicative of assembly structure naturally having no associated clustering measure, when considering a lone assembly. As such it would be expected that network 3.1 and 5 would have clustering coefficient measures, as they are collections that contain many assemblies. In network 2 the highest clustering coefficient value belongs to the parts named ‘LValve’ and ‘RValve’, which are the two parts connected by an edge noting their ‘mirror’ relationship. In network 3.1 these nodes are also recorded as having the highest value for clustering coefficient. This would appear to be due to the multimodal edge that connects them, despite the many separate components that network 3.1 contains.

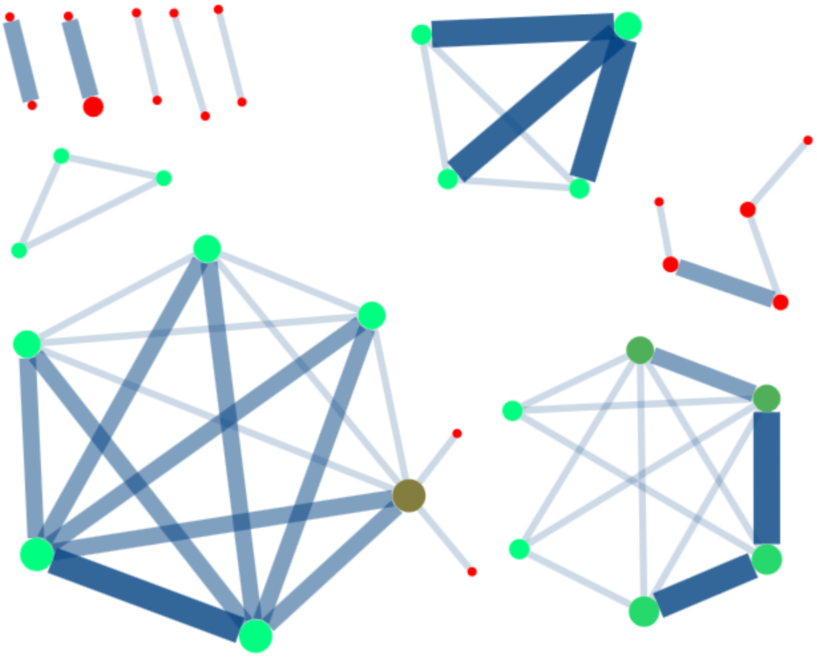


Figure 4.12: Network 4 with clustering coefficient mapped to node colour

Network 4 has a higher average clustering coefficient than networks 1, 2 and 3.1, and 9 nodes that are measured as having the maximum clustering coefficient, suggesting that it does have a small world structure. In figure 4.12 the clustering coefficient of each node has been mapped to the node colour, with red showing a low clustering coefficient and green indicating a high value. From figure 4.12 it is plain to see that the nodes that are most connected are those which have the highest clustering coefficients, however there are no edges linking the highly connected clusters together. From this it could be concluded that linking 3D designs by shape similarity results in a small world network structure being formed. This could be of importance as shape similarity work continues.

Also of interest is the similarity between the clustering coefficients found for network 3.1 and network 5. Despite the significant differences of size and density of these networks, the CAD assembly data they model has resulted in the same low measure for this metric. From this it could be concluded that CAD data, when modelled by assembly structure in a network, does not conform to the small world model. Telesford *et al.*'s measure could be used to assess this further. It is also notable that the similarity between the Edinburgh Benchmark and the real world CAD data might verify the structure of the data in the benchmark.

4.7 Hubs and Connectors

Barabási highlighted the significance of hub nodes; nodes which have more edges connecting them to other nodes than would be expected within a network. Despite Barabási's work, Watts disagreed that hubs are pivotal, after his 2001 experiment concluded few traversed paths included the crucial 'hub' nodes Barabási cited as important. In social science hubs are often interpreted as well-connected people, but in mechanical engineering design networks a well-connected node is not so easily classified and it depends upon the relationship between CAD files that is being modelled. In an industrial situation, where a collection was modelled by assembly structure, a well-connected node could indicate a well used part and as such show where efforts in optimisation should be focused. In an educational setting, where shape similarity relationships were modelled, a well-connected node would indicate a model that was similar to a lot of other work and as such could be of concern. Table 4.5 shows the degree measurements for networks 1-5.

	Network 1		Network 2		Network 3.1	Network 4	Network 5	
Degree	In-	Out-	In-	Out-			In-	Out-
Minimum	0	0	0	0	1	1	0	0
Maximum	1	5	4	8	40	7	534	151
Average	0.89	0.89	1.39	1.39	2.25	2.94	4.95	4.95

Table 4.5: Degree metrics for each network

An initial note on comparing the degree measurements of networks 1-5 is that these metrics are determined by the graph type, while being node specific. Networks 3.1 and 4 are undirected and as such have only one measure for degree, while networks 1, 2 and 5 are directed so have measures for both in-degree and out-degree. The minimum degree for an undirected network must be 1; if the minimum degree for an undirected network was 0 it would indicate that there were unconnected nodes in the network. In network 3.1 this would show a part that was not used, so could be a useful check when modelling CAD assemblies. In network 4 it would be illogical, as the relationship modelled is similarity, and so a minimum degree of 0 would indicate an error.

In networks 1, 2 and 5 a node with a high in-degree would indicate a part that was used many times, while a low in-degree would show a part was not commonly used. A high out-degree would denote an assembly that contained many parts and a low out-degree would indicate an assembly that contained few parts. A node with an in-degree of 0 would indicate a top level assembly, while a node with an out-degree of 0 would be a singular CAD component. These interpretations are similar for network 3.1, where a high degree would indicate high use of a part of an assembly, however as the network is undirected it is unclear which parts are top level assemblies and which are low level components from these measures.

The average degree indicates whether most nodes have a high or a low degree. In all cases for these networks the average node has a degree of less than 5, showing all CAD data modelled is connected to fewer than 5 other nodes directly. In a large collection, such as network 5, this could indicate low use of many CAD parts and could be useful in exploring which CAD parts should be improved and used more, or which could be removed from a company's collection as they are less important. The overall low average degree suggests that few of these parts would be hubs, however in an industrial setting it could be most profitable to focus on optimising the low level components which are often used, shown by a high in-degree.

The maximum degree for nodes in network 4 indicated how many times CAD files submitted by a student have been identified as geometrically similar to another's and as such could play a key part in identifying work that is unoriginal. If a node within network 4 is highly connected, determined by degree, it is more likely that the student's work is similar to others', than where a node has few connections. This measure could be used as part of an analysis of students' work, allowing a marker to locate copied work within a large collection of submitted designs.

With regards to search capabilities, CAD 'hub' nodes would play an interesting role in search, acting as connectors between many parts and potentially being part of commonly traversed search paths. Watts' conclusions, which seek to disprove Barabási's theory, are not applicable here, because Watts focused on networks where nodes were people who could actively take part in deciding which links to use (which edges to traverse). As CAD files are inanimate, they would play no role in the search path traversed and as such it may be illogical to attach a measure of connectivity to a node in a network like networks 4 and 5. By determining the place of a node within the network, using a metric such as betweenness centrality, and comparing it to how often it is used in a ‘contains’ relationship, by using its in-degree, a way to identify these key parts could be found and this could be instrumental in highlighting which parts are ‘hubs’ in these networks. These ‘hub’ nodes could be of more importance in a network where CAD files are linked by shape for search.

4.8 Weak Ties

In his 1973 paper Granovetter concluded that weak ties were more useful socially than strong ties, when looking at social networks formed by people. He suggested that these play a key role in forming social networks, where someone may be more likely to be given a job opportunity through a socially weak tie than a strong one. For the CAD files modelled in networks 1, 4 and 5 the strength of the ‘tie’ between two nodes could be considered as measured by the associated edge weight, which in these networks is visualised as edge width and opacity.

Figures 4.2 and 4.7 show networks 1 and 5, CAD data modelled by assembly structure, where a heavy edge between two nodes indicates a part that is used several times within a containing assembly. In network 1 there is only one such edge, showing that the CAD part represented by the node in the bottom right-hand corner of the image is contained more time within the sub-assembly one level up from it than the other parts on the same level. In figure 4.7, there are many heavy edges shown within network 5 as well as many thin, transparent edges showing weak links, representing CAD files that are not used many times. When considering the strength of these links, a weak link could indicate a lesser used part within a large collection and as such indicate parts requiring attention or improvement, while strong links could indicate those parts that would be best to concentrate on for optimisation, due to their regular use.

Figure 4.6 shows network 4, where the relationship modelled is the geometric similarities between students’ work. When considering the links in this network a heavy edge, representing a strong link, could be indicative of high levels of similarity between students’ work. Using this measure, along with others such as degree, could be a way to locate work that was unoriginal.

Socially weak ties provide people with new information and opportunities and, according to Granovetter, are useful. It was considered, when applying network theory to mechanical engineering design, whether these weak ties would provide similarly advantageous situations and information. In a design scenario a weak tie between two CAD files could provide a designer with the link to a much needed design solution. In network 5 a weak tie could be key in connecting different assembly structures that would otherwise be independent of each other. In network 4 a weak tie could be indicative of important similarities between students' work, as it is possible a student may have copied a single file from another. However in networks 1, 4 and 5 it can be concluded the strong links, represented by wide edges, are more in useful and important in this analysis.

4.9 Layout

Each network presented in this chapter has been carefully arranged to provide an optimised image for analysis, however there are many differing types of layout. Without being given further information or a key, different layouts can communicate differently and highlight different characteristics of a network. It is vital to consider the impact that a visualisation has on understanding these structures. Network 4 provided various opportunities for examining layout options as the relatively small number of nodes and edges, alongside the various individual clusters it contains, necessitated care being taken when attempting to create an effective visualisation.

The various types of layout available can be categorised as

- Force-directed
- Circular
- Arc
- Grid
- Tree

and this work made use of those available in NodeXL. Figures 4.13 to 4.18 show six different network layout options.

Figure 4.13 and figure 4.14 show two commonly used layout options; the Harel-Koren Fast Multiscale layout and the Fruchterman-Reingold force-based layout respectively. The Harel-Koren Fast Multiscale algorithm is a finely tuned algorithm for effectively laying out networks, but can be too computationally intensive for large networks. A force-directed layout, such as that shown in figure 4.14, is considered to work like a set

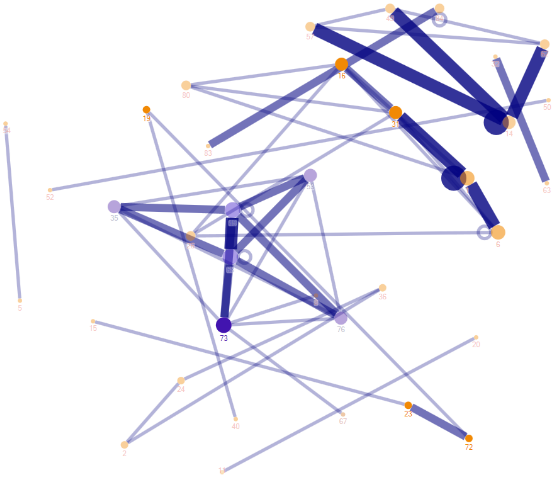


Figure 4.13: Network 4: Harel-Koren Fast Multiscale layout

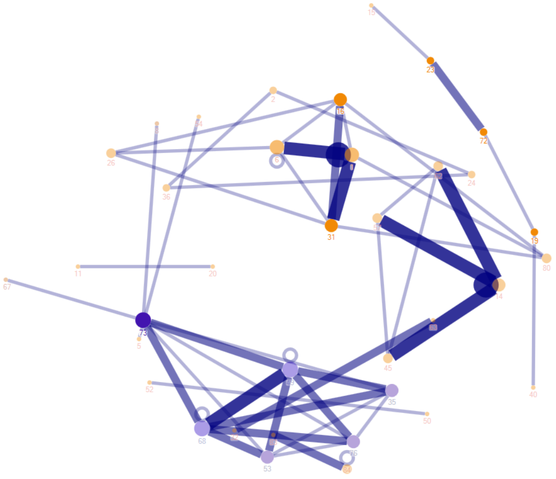


Figure 4.14: Network 4: Fruchterman-Reingold layout

of springs, where different measurements are prioritised and based upon these the nodes and edges are arranged automatically and are located according to Hooke's law; linked nodes attract each other, while non-linked nodes are pushed apart [71]. They can be very useful when displaying large scale networks, though they can be computationally expensive. For network 4 neither layout was considered effective as the overlapping edges did not provide a clear diagram of the data.

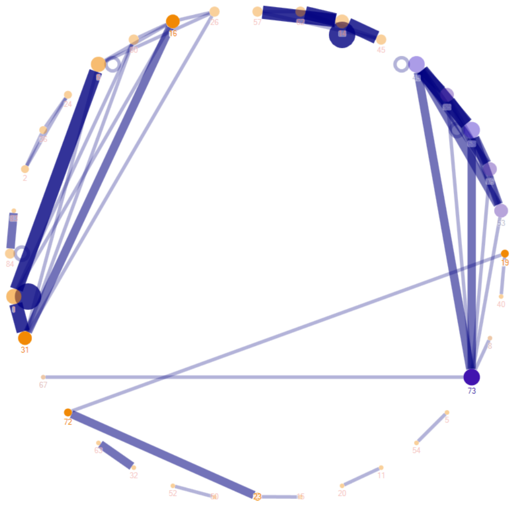


Figure 4.15: Network 4: Circle layout

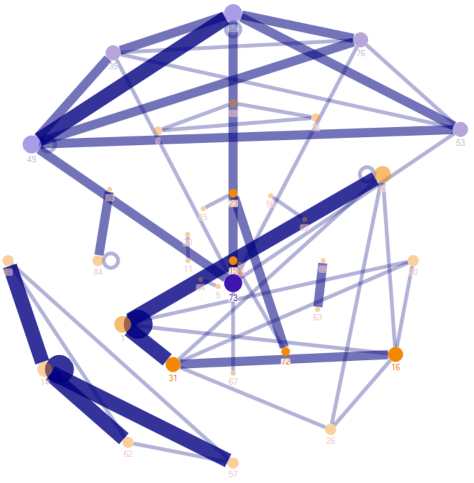


Figure 4.16: Network 4: Spiral layout

Figure 4.15 shows network 4 in a circle layout while figure 4.16 shows the same data in a spiral layout. These two are examples of shape based layouts, specifically circular and arc based. While the circle layout was effectively used to illustrate the differences between networks 3.2 and 3.3 in figures 4.10 and 4.11, it is less effective here. It shows that network 4 has several separate clusters, but does not represent the structure of the data well. Similarly figure 4.16 using the spiral layout produces a complicated diagram, with little clarity.

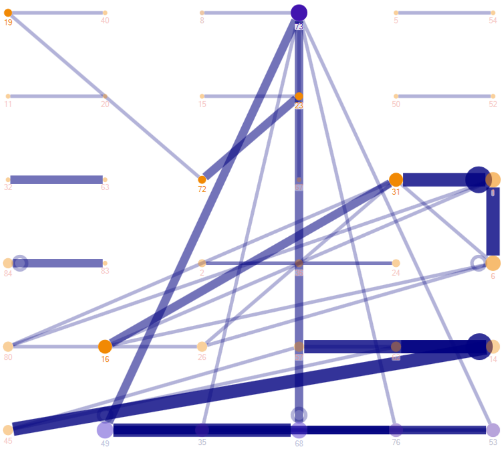


Figure 4.17: Network 4: Grid layout

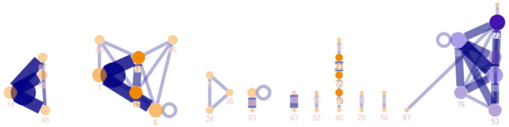


Figure 4.18: Network 4: Sugiyama layout

Figure 4.17 shows network 4 arranged using the grid layout, where the some of the separate clusters can be more easily identified. The rigid structure of this layout makes the relationships modelled appear more straightforward by use of straight lines, however as no weighting is applied to the nodes, some that are closely linked are far from each other, located on opposite sides of the diagram. Because of this it is difficult to assess the clusters or clearly see how they are linked. Figure 4.18 shows network 4 arranged using the Sugiyama layout, which is an ordered tree layout. Here the individual clusters contained within the network are separated and clearly visible, but this layout does not provide a clear image of their structure. This layout was used to effectively model network 1, shown in figure 4.2, as it reflects a traditional tree diagram commonly used to describe assemblies. However, as network 4 does not model an assembly structure it is unsurprising that this layout is not effective. This was also the case for network 2, illustrated in figure 4.3.

After exploring these layout options it was decided that the network 4 would be arranged by hand. This was found to produce a clearer image, shown in figure 4.6. However this is not the case for all networks presented in this chapter and some made effective use of the layouts provided in NodeXL. As this work continues, layout options will be carefully considered and chosen to optimise the visualisation of networks.

4.10 Discussion

In this chapter 5 novel networks have been presented and analysed with respect to key discoveries within network theory. The images shown here are also novel, as networks of CAD data have not been built or presented visually outside of Mill's research group.

When analysing the metrics of networks 1-5, it was found that several highlighted interesting characteristics. Assessing the metrics alongside a clear visualisation of the data provided clear information on the size of a network as well as its structure. An original interpretation for these values, when calculated from networks of CAD data, was presented in table 4.2 and random CAD networks were also analysed, with network 3.2 and 3.3 being presented in sections 4.4. It was concluded that random networks of CAD data are meaningless as they do not represent real mechanical engineering design structures.

The size and density of networks 1-5 were assessed and compared throughout this chapter and determined to be indicative of characteristics about the modelled data. It was found that CAD networks do not conform to Milgram's accepted 'six degrees of separation' theory and CAD assembly networks also do not conform to Watt's small world theory, while network 4, modelling the similarity of students' CAD work, may conform to it. The hubs that Barabási argued were important in networks can be seen as important in CAD networks also, not due to the influence they have but because of the data they represent. Granovetter's weak ties theory, however, was not seen as pertinent when considering the links between different CAD data.

Visualisation was briefly assessed and various visualisations of networks were presented. While using these layout options is not novel, applying them to networks modelling collections of CAD data is unique. The influence a layout has over how a network is perceived is of vital importance and this work will seek to present clear and effective visual representations of all the networks presented.

Key conclusions from this work to consider taking forward include the important metrics found for network 4. These include degree and edge weight, which seem to indicate important information about the similarity between students' work in network 4.

Another area for further consideration is the mode of a network. Several of the networks here were not unimodal and it can be seen that collections of CAD data do not easily lend themselves to unimodal representation. Though for the purpose of this initial, simple analysis some networks were discussed as if they were unimodal, due to the equivalence of CAD files, naturally the data structure of assemblies is not unimodal and as such would be more accurately considered bimodal and modelled accordingly. Assembly data can be expanded and displayed as multimodal by adding additional relationships and files, as network 3.1 illustrated.

When considering multimodal networks, large collections of company PLM data could be used and it would be possible to include nodes representing other areas of the product lifecycle, intrinsically linked to the CAD models. This could be done by taking the data held within company PLM systems and assessing the relationships between CAD, orders and customers, and creating a network of this data. In this way a multimodal network could be made by expanding network 5. If additional data, linked to the CAD collection was added to the network, it would be possible to create a network where nodes could represent CAD components, sub-assemblies and assemblies as well as orders placed for them by customers, each of whom would be a node and an edge would represent an order placed. While this network may quickly become large and complex, it could be highly valuable due to the inherent worth of this kind of data, by providing commercially useful insights.

4.10.1 Further work

From the networks presented here it is clear there are some key areas which provide opportunity for further analysis.

Considering CAD in education, network 4 was created to model the geometric similarity between students' CAD work. It was found that degree metrics and edge weight could be indicative of strong links between students 'folders' of work and these strong links could show deep similarities, and even indicate unoriginal or copied work. Investigation into this line of assessment therefore seems highly promising.

Considering CAD in an industrial setting, the intrinsic value of the data warrants further investigation through multimodal networks. Large networks could be built to model the data structure present in company PLM systems, showing the whole of a product life cycle and effectively encompassing a bill of materials.

Networks may also provide advanced search opportunities for geometric matching analysis. This work could begin by assessing Milgram's theory, investigating if any two models could be reached within 6 steps within a CAD collection. Also a measure for searchability could be defined, using the available metrics to produce a characteristic equation describing a CAD collection.

Chapter 5

Unimodal Networks

Investigation of networks built from CAD data in an educational setting: Analysis of shape similarity technology and network theory to effectively identify plagiarised work

In the previous chapter an investigation into network theory with regards to CAD data was begun. Based upon the background and motivations presented in chapter 2, chapter 4 presented five novel networks build from CAD data and analysed them according to key advances in network science. From both chapters the value of CAD data has been asserted and the area of education highlighted. In particular, the work done in sections 4.7 and 4.8 called attention to the possibilities that particular metrics might allow for networks of CAD education data. The discussion on mode presented in section 4.2, and elsewhere in chapter 4, conveyed the complexity and inherent issues involved when working with bimodal and multimodal networks and as such it was decided to create unimodal networks, similar to network 4 presented in section 4.2, to investigate CAD collections in education.

In this chapter a novel method for creating network diagrams of collections of submitted CAD data from university classes is presented in subsection 5.2.1 and a methodology for discovering unoriginal work is proposed in subsection 5.3.3. This method is simulated several times, before being assessed using real world data in section 5.6. This chapter concludes the proposed method is robust and reliable when identifying unoriginal work and recommends areas for further work.

For this investigation results from ShapeSpace technology were used, making it possible to accurately and reliably identify geometrically similar or identical 3D CAD models within a collection of parts. These results were then analysed to show the links between the students in a network graph using NodeXL, and the graphs presented are screen shots from this software.

5.1 Introduction

After an initial analysis of network theory applied to CAD collections, it became clear that there was a basis for an interesting analysis exploring collections of CAD models in an educational setting. The different types of networks, which were presented in chapter 4, each represented differing collections of CAD models. Some of them were built with one type of node and one type of edge (Networks 1, 4), others were built with nodes and edges that represented a variety of components (Network 2, 3.1 and 5).

It was determined that a key area for this work would be unimodal networks. In chapter 4 different key elements of several networks were explored, including the various metrics. The most common metrics are defined with respect to unimodal networks and become difficult to interpret when more than one type of relationship is modelled. This was a principal consideration in choosing to further investigate unimodal networks.

Modelling collections of CAD files linked by shape would provide a unimodal network for investigation where the nodes and edges were static. Exploring these simple data structures provides a chance to further explore their structure and metrics, and could provide new insights into the collections of CAD models they represent. It was also determined that investigating networks with more than one type of node and more than one type of edge would be of interest, and this work is presented in chapter 6.

In chapter 4, models of CAD collections in educational settings were identified as the most promising area for advancement for several reasons. With many years of data available through the university database, it was evident that it would be possible to build and compare several networks from classes of different years. The wealth of data available also emphasised how there are large collections of CAD models in this educational setting. This volume of CAD data is commonplace in educational institutes where students are taught to design and use specialist software, including in universities and some schools. Therefore it is of high value to educators and worth exploration.

It was suggested that building a network of students' design files would be a useful way to compare the similarity of their work. This kind of network would be original and novel, because nowhere in the literature has the similarity of 3D CAD models been used to build a network. Modelling the students as linked by their work, rather than linking them socially, would provide a different insight into their interactions and could even be an indicator of duplicated work, highlighting students who worked closely together or even those who had copied another's work.

This research was determined to be of potential value because uncovering new information would assist educators and provide insights into collections of students' CAD models. With many schools and universities teaching 3D modelling and CAD, students produce sizeable portfolios of work. Assignments are set to determine and monitor a student's ability and understanding, both of meeting design briefs and the software they are using. This generates a significant amount of work for educators to assess and mark. These types of hand-ins often comprise written reports, digital technical drawings, animation files and 3D CAD files. Each student turns in a number of CAD files resulting in staff having a large collection to look through and the complexity and volume makes it nearly impossible for a teacher, lecturer or marker to notice plagiarism in 3D hand-ins. In universities where class sizes are large, it would be unrealistic to expect a marker to identify copied files.

It was suggested that making use of ShapeSpace's software to produce a similarity analysis of the 3D CAD files that a selected class had handed in could provide information that would assist marking of these types of assignments and possibly help determine if files had been copied. This analysis would have to begin with a duplicate and similarity analysis of the 3D CAD parts. Duplicate analysis is a widely investigated topic, see subsection 2.5.1; however for this investigation it is the use of the analysis results that is of interest.

Considerable work has been done on improving shape similarity and duplicate search technology and very accurate results can now be computed to reveal geometrically similar or indeed identical parts. However, as discussed in subsection 2.5.1, these results are often of no value unless further investigation is done. While geometric duplicates are now easily identifiable in a range of ways, there is still a struggle to identify semantic duplication in 3D work. The difference between geometric duplicates and semantic duplicates also remains under-discussed and under-investigated, with currently limited uses. Another novel aspect to this proposed work would be that it begins to make use of those results in this engineering application; building a network of how collections of 3D work are similar by shape would be a creative and original use for these similarity analysis results.

In the initial shape similarity analysis of the data collection, some duplication may be identified. When searching for or identifying duplicated CAD models, the type of duplication must be considered. It is true that within a network there will be both geometric duplicates and semantic duplicates. In a large part collection geometric duplicates are not necessarily engineering duplicates; two geometrically identical part made of different materials may both be valid. A collection of CAD parts from an educational setting such as those explored in this work will contain many semantically similar parts, as all students will have been designing something to meet the same brief. This should not mean, however, that their designs are geometrically identical when compared.

There may be legitimate situations where students' work is geometrically identical, for example when a design brief sets certain parameters or when designing something to be part of an existing system. As data is used to build networks this investigation will seek to identify if admissible geometric similarity is detected and causes issues during the originality analysis.

There have been several motivations for this work. Firstly is the limited availability of software and lack of an established method at present for detecting the academic misconduct that is plagiarism in non-text based, 3D CAD assignments. There are many different assessment tools which can be used to detect plagiarism in written assessments, one popular example of this being turnitin which is used within the University of Edinburgh and by many other educational establishments. This kind of software to aid detection of academic misconduct is not only widely used, but also is a well established, and has become an excepted part of academic culture within subjects that require written assessment. There is currently no such aid for educators who practice non-text based subjects, which require visual, verbal or alternative file assignments and examinations.

Secondly, plagiarism is becoming a more prevalent problem as computer technology use increases. The increased availability of free information online is an advantage within modern education, but also causes problems; some students not only share work online but also use information without disclosing its source. Students are taught about plagiarism and how to avoid it, by referencing work correctly, but there are often cases where students seem ignorant of the correct procedures, though deliberate plagiarism also occurs. This is not only a problem in text-based subjects, but affects every area of education. There are even reports of students now spending time working out how to get around detection techniques or claiming they have a right to plagiarise material, rather than simply learning the required material for a course or qualification.

Lastly, this chapter will attempt to discover if shape similarity can usefully assist in detecting the originality of students' work. Shape similarity software is advanced, but has not been put to use in this context before. When combined with network theory and the metrics it provides, shape similarity may hold the key to identifying unoriginal files and therefore assist a marker as they seek to assess many students' work. With this in mind, this work begins by analysing the work of a historical class from the University of Edinburgh.

5.1.1 Aims of unimodal network exploration

For the initial investigation of this data it was determined that a network should be built from the hand-in data of one previous class at the University of Edinburgh. Initially the 3D CAD files were to be collected and compared by shape, then a network would be built, where the nodes would represent a student's body of work and the edges would represent similarity links between students' work.

The initial aims of this chapter's investigation were set to:

- Assess the usefulness of modelling CAD parts when linked by shape.
- Assess this use of network theory in describing a collection of CAD models from one class's submissions.
- Assess this use of network theory in identifying engineering duplicates, not just geometrical identical parts.
- Assess this use of network theory in assisting with marking large collections of work.
- Assess this use of network theory in identifying copied work, i.e.: plagiarism.

5.1.2 Proposed social dimension research

An additional proposal was made with the aim of strengthening the analysis of the network and to assist in achieving the aims outlined in subsection 5.1.1. It was proposed that data could be collected about the social grouping of the students and added or compared to the data gained from the shape similarity analysis, to build a network which would provide a comprehensive overview of the students' interactions.

This information could be collected in several ways. The first way would be to use data on the groups to which students were assigned. A second way would use readily available statistics from the university course web page and the course Facebook group. Connections between the students could be mapped using the data on who was a member of the Facebook group, who was actively participating in discussion via the group, who was visiting the course web page, as well as who was engaging via discussion and the web Q&A section. A third proposed way was to survey the students and require it to be submitted along with coursework.

The questionnaire would seek to track friendships and work groups, asking each student to rank each course mate on a scale including 'I don't know this person', 'acquaintances' and 'close friends'. Another proposed format for this questionnaire was for students to simply write down the names of those they knew in the class. Either of these questionnaires could be completed at the start and end of each semester to map the change in friendship groups. This would also provide interesting information on how the groups assigned for courses affected the friendship groups of undergraduate engineers. This could be combined with formally recorded group membership in the class. Thus the strength of relationships between everyone in the class could be measured. Tracking the progression of relationships and use of university sites over the years would build an interesting map that could be compared to historic class data and again highlight possible plagiarism links.

However, as much of the available data was historic, it would not be possible to gather information from the course website or use a questionnaire to collect statistic about friendship circles. While this could be overcome by assessing the data of a current class, concerns were raised about the questionnaire. It would be unethical to expect the students to give time to completing the questionnaire without first explaining the reasons for it, but discussing the goal of the questionnaire may result in students not giving honest answers. Also discussing the plagiarism investigation with the students may shift their focus during the course, meaning they act differently as they may be overly conscious of the aims of this research. It would also be inappropriate to discuss this investigation with the students in case it makes them feel untrusted or as though they were being observed and monitored in ways they aren't in other courses. Also as a distraction from the education being given, this would potentially hinder accurate or reliable results, thus negating this additional proposal.

For these and other reasons this proposal was rejected. One reason for rejecting the proposal was that learning more about the social network of the students, while both interesting and a novel way to compare work, would not provide any truly meaningful extra data for this investigation of a unimodal network, or provide much more to support the stated aims in subsection 5.1.1.

5.2 Network 6.1

Initial Network Model

This section reports on the first unimodal network to undergo the analysis technique using ShapeSpace's program and the open-source network software discussed in section 3.2. The network was created from the anonymised files of one class's submissions for the 3rd year CAD course at the University of Edinburgh, referred to as network 6.1

5.2.1 General analysis method

Building a network, which is then visualised, takes several steps involving different software and programs. These steps are discussed below and are entirely novel and specific to this work. All the programs used are discussed in chapter 3.

To create the network a folder is created, containing the submitted files from every student on the course. This folder is analysed with ShapeSpace's similarity software, with each file first being indexed before a duplicate analysis is run.

Once the duplicate analysis had been performed, a tab-delimited text file is produced and this contains a list of the groups of duplicates that have been identified. This can be turned into a list of one to one relationships, which can then be input to NodeXL.

The network is subsequently built containing nodes which represent an individual student's file collection, and edges representing the links between similar or duplicated files.

5.2.2 Network 6.1 inspection

For this initial network graph the sample collection of one class's submissions were analysed with ShapeSpace software and the results entered into NodeXL. What follows is the novel process used to create a useful network graph from these raw results.

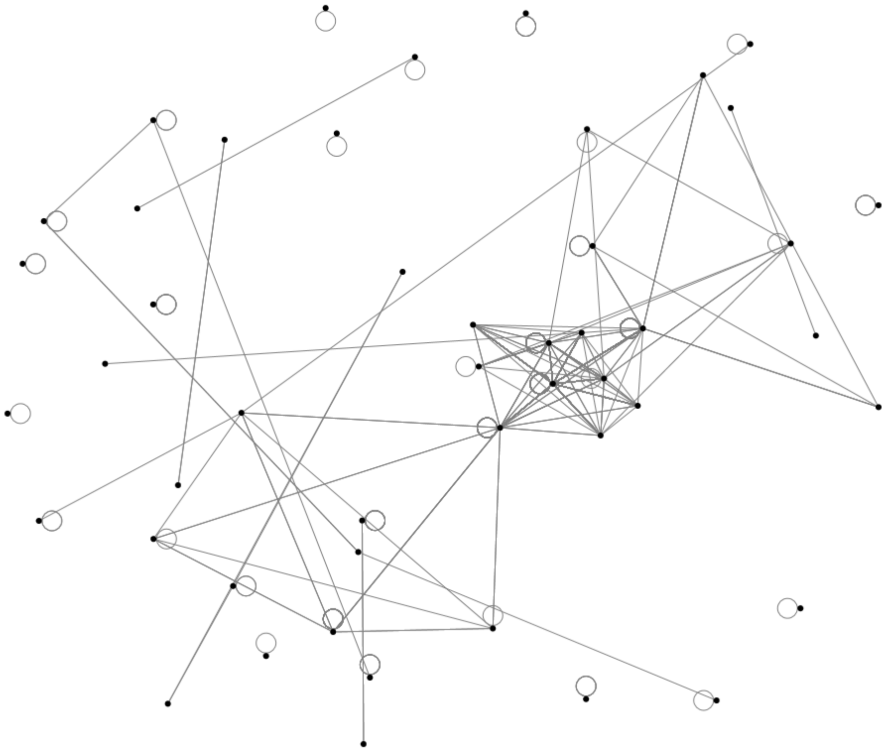


Figure 5.1: Network 6.1 produced from one class's CAD course submissions

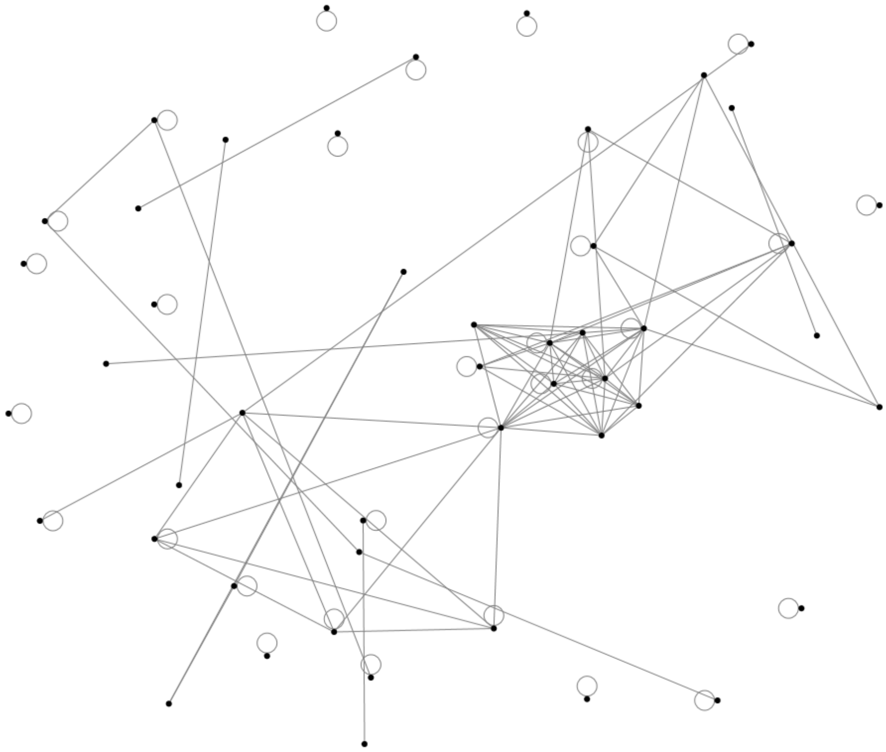


Figure 5.2: Network 6.1 with duplicate edges merged

In figure 5.1 the initial graph produced from the raw data is shown. NodeXL uses the Fruchterman-Reingold layout for the data as default. It is evident from figures 5.1 and 5.2 that this is not the most favourable or clear layout for this data. The raw data shown in figure 5.1 had many duplicate edges, and these were counted and merged within NodeXL. The duplicate edges count was then transformed into a value for edge weight, giving more graph metrics to work with. While figure 5.1 shows the raw data as initially displayed, figure 5.2 shows the result of merging the duplicate edges.

The images are only subtly different, which highlights how duplicate edges can cause problems for understanding network diagrams. If this network were to contain multiple edges, instead of these being consolidated into a property of one edge, the calculations for the other graph metrics would be affected. One example of this would be the calculated degree of a vertex, which would be higher if duplicate edges were not merged. For this data it makes sense to merge the duplicate edges because the number of connections a student has to other students is of interest, rather than how many connections they have overall.

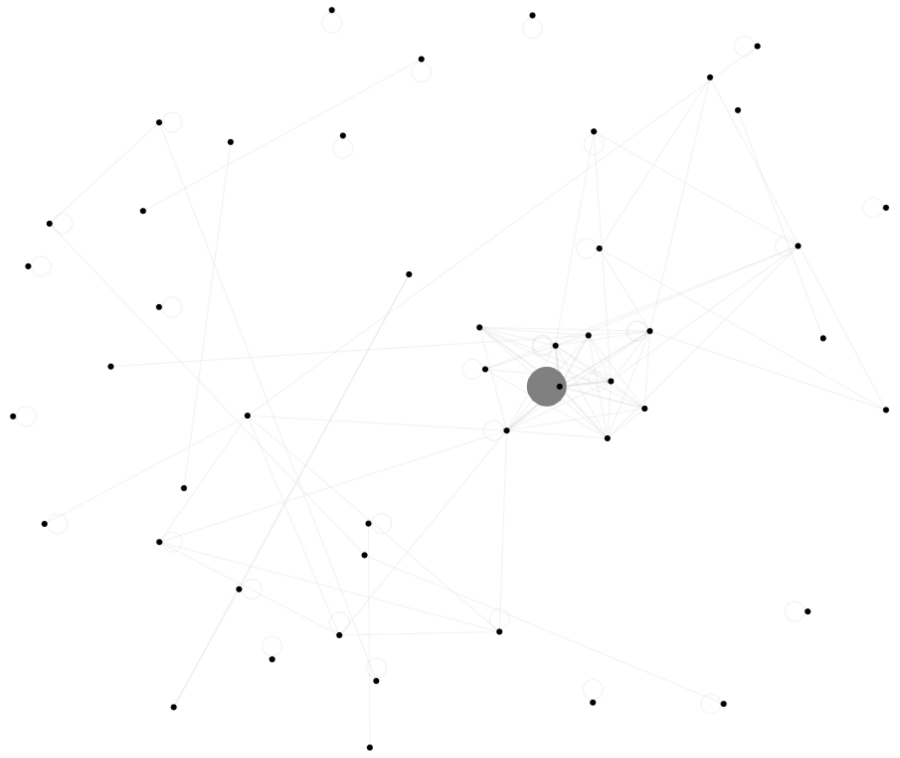


Figure 5.3: Network 6.1 with initial mapping of metrics to visual properties

Once the duplicate edges were counted and merged, the other network metrics were calculated and assigned as visual properties to the nodes and edges in the network.

Figure 5.3 shows the initial results of mapping edge weight to edge widths and edge opacity. This image is of little to no use as there is one edge within the network that has such a heavy weight that it makes all the others appear extremely small and transparent. To solve this problem, NodeXL has a function called ‘ignore outliers’. This function allows a maximum value to be set for a metric, in this case edge weight, and any edges with a high ‘outlier’ weight are set to the maximum value assigned.

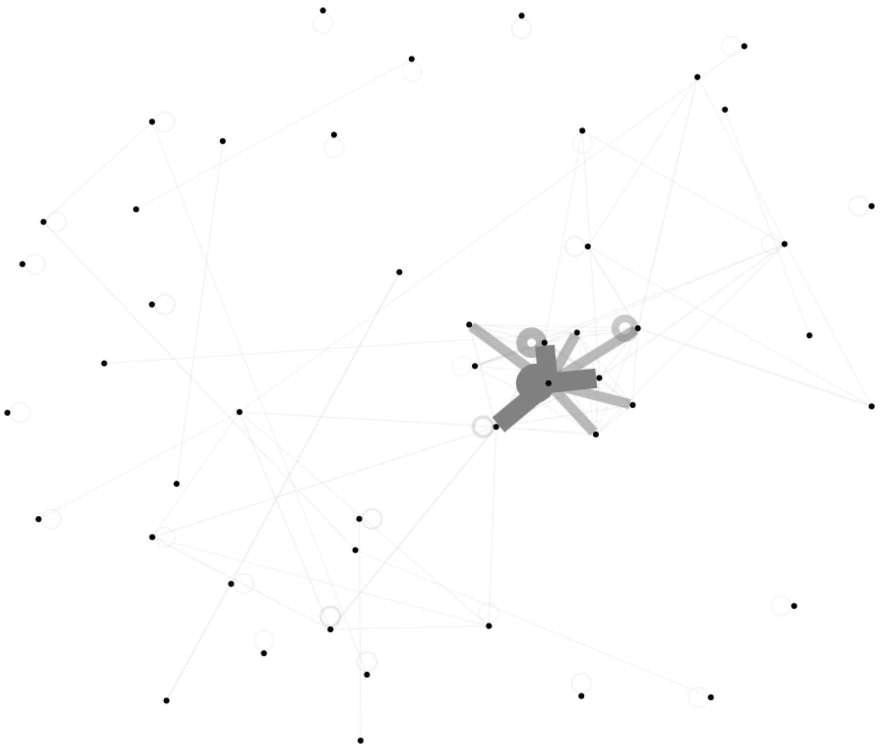


Figure 5.4: Network 6.1 once ‘ignore outliers’ function has been activated

Figure 5.4 shows the effect that the ‘ignore outliers’ function has on the network visualisation. The edge that was the heaviest is still clearly shown, but others are now also highlighted. It is worth remembering that while edge weight and opacity now represent the number of links between two connected nodes, this is often interpreted as how ‘strong’ the link is between those two nodes, here two students. This is a rational interpretation, as the more connections shown by multiple edges, before they were merged, between two nodes do indicate a stronger connection between them than two nodes that have fewer edges linking them.

Once the duplicate edges had been counted and merged, it was possible to calculate the graph metrics and map those to node and edge properties. This again improves the network visualisation. From the mapping shown in figure 5.5 it can be seen that there

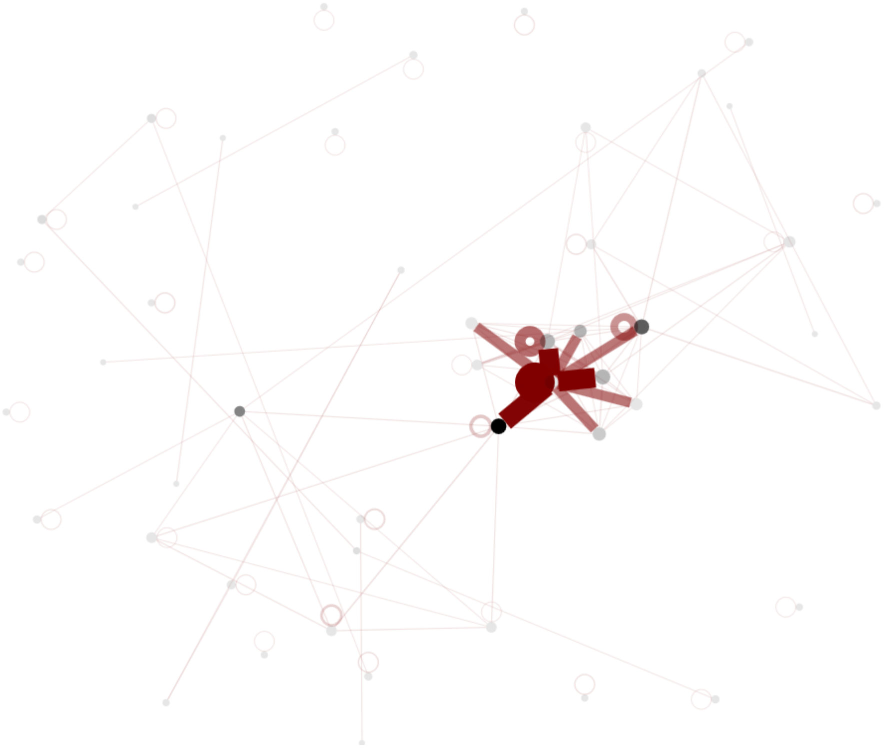


Figure 5.5: Network 6.1 with first visualisation of metrics

are some nodes which now stand out within the network. The red colour assigned to the edges was chosen to assist with clear visualisation.

The assigned metrics are:

- Node degree was mapped to the node size
- The betweenness centrality was mapped to the node opacity
- The edge weight was mapped to both the edge size and opacity, with outliers ignored.

These were chosen as they were deemed to be the most important metrics for this network.

While figure 5.5 shows the first attempt at visualising the calculated metrics, it is in the default layout for NodeXL, the Fruchterman-Reingold layout. Figure 5.6 illustrates an attempt to use a different layout option, the ordered circle, rather than the default Fruchterman-Reingold. Figure 5.5 has been included in this work to illustrate that the network was rearranged using numerous automatic layouts provided within NodeXL. However while forced or ordered layouts can reveal different properties of networks or

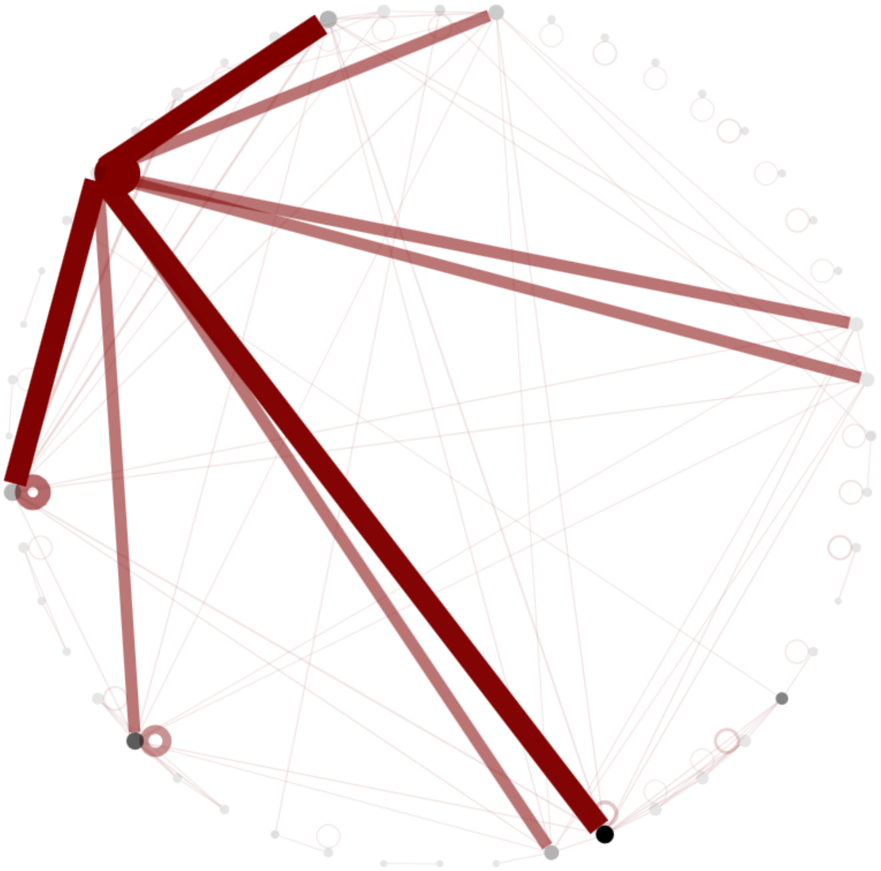


Figure 5.6: Example of alternative layout for network 6.1

certain features, it was not found that any automatic layout gave a clear visualisation for this network.

As no automatic layout seemed to effectively convey the structure of the data, the network was laid out by hand, allowing for a visually clear appearance (see figure 5.7). From this it became clear that there were several key nodes to investigate. To further facilitate this investigation additional metrics were mapped as visual properties to the graph, shown in figure 5.8.

The changes in the graph visualisation between figures 5.7 and 5.8 were chosen to further assist in visual inspection of the data. Different colours used were chosen to show a clear difference between visualisations as this data investigation progressed. In addition to this other metrics were mapped as visual properties onto the graph. The assigned metrics were now:

- Node degree was mapped to the node size
- The betweenness centrality was mapped to the node opacity

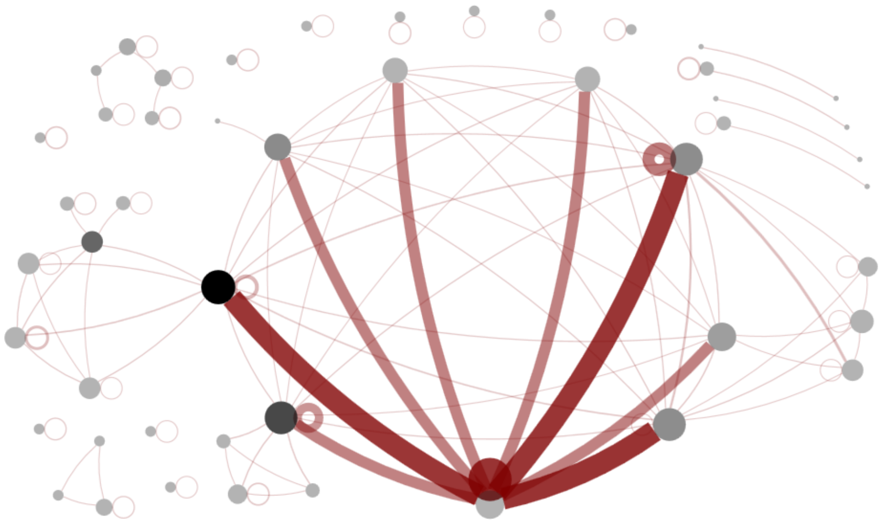


Figure 5.7: Network 6.1 laid out by hand

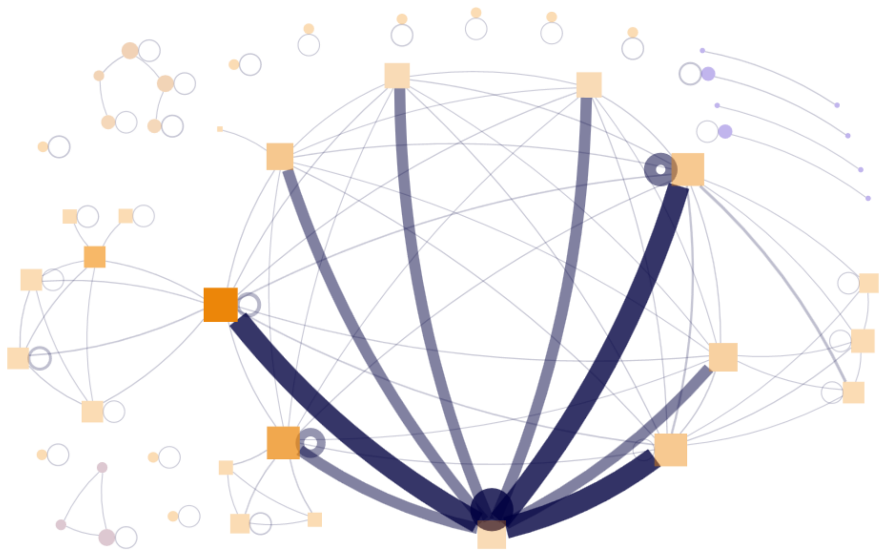


Figure 5.8: Network 6.1 with further visualisation of metrics

- The closeness centrality was mapped to the node colour
- The eigenvector centrality was mapped to the node shape
- The edge weight was mapped to both the edge size and opacity, with outliers ignored.

These metrics were mapped in this fashion to facilitate further investigation of the data. By mapping metrics including closeness and eigenvector centrality, their importance could be determined visually.

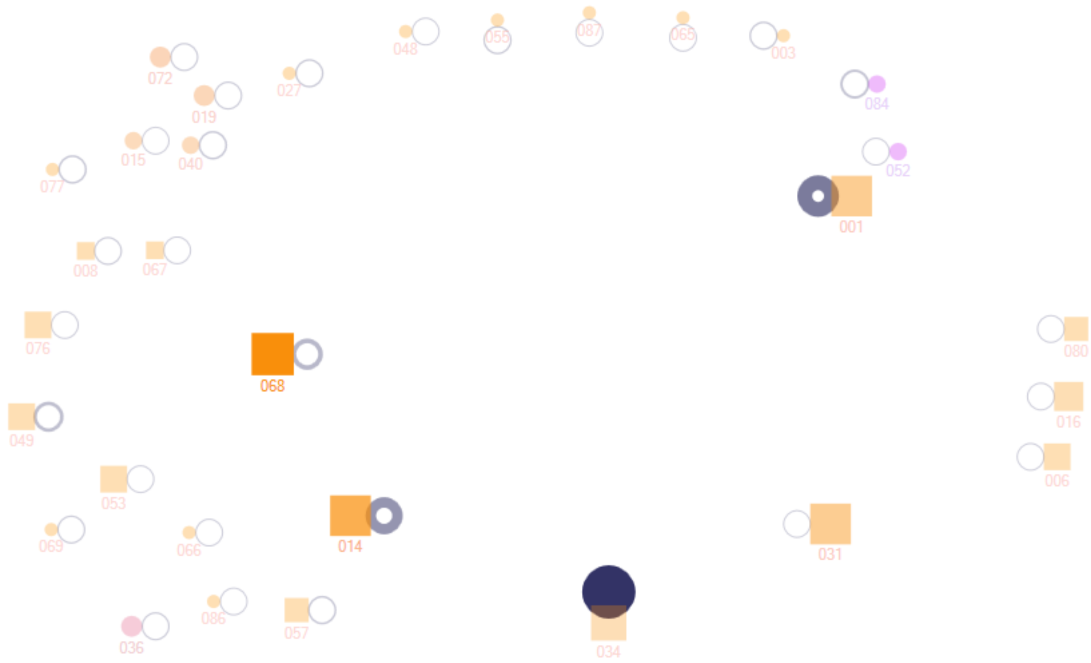


Figure 5.9: Self-loops from network 6.1 shown in isolation

An important feature to note in figure 5.8 are the nodes that are linked only to themselves which stand apart from the others, shown at the top of the diagram. Figure 5.9 shows the network graph reconfigured to show only nodes that are linked to themselves. An edge returning to the same node it comes from is termed a self-loop and reveals previously unknown information. Self-loops indicate several things depending on the data they represent but in this case it most likely illustrates that a student has work within their directory which is similar to other work in their own directory.

It is also important to note that the large outlier, highlighted in figure 5.3, was a self-loop, and this is further discussed in subsection 5.2.3. For simplicity the self-loops were removed from the visualisation. Also visible are the labels now included in the graph, to show the students anonymised identity, allowing the nodes to be identified visually. Notably at this point it seemed reasonable to exclude self-loops from the network diagram, because when a student's work was linked to itself it was not a cause for concern; a link from a student's work back to the same work could not indicate plagiarism. Later in this investigation however, it became important to retain this information and self-loops were included in visualisation (discussed in subsection 5.2.3).

Figure 5.10 shows the final network graph of the student submissions. It has been refined, and now clearly illustrates the graph metrics simply.

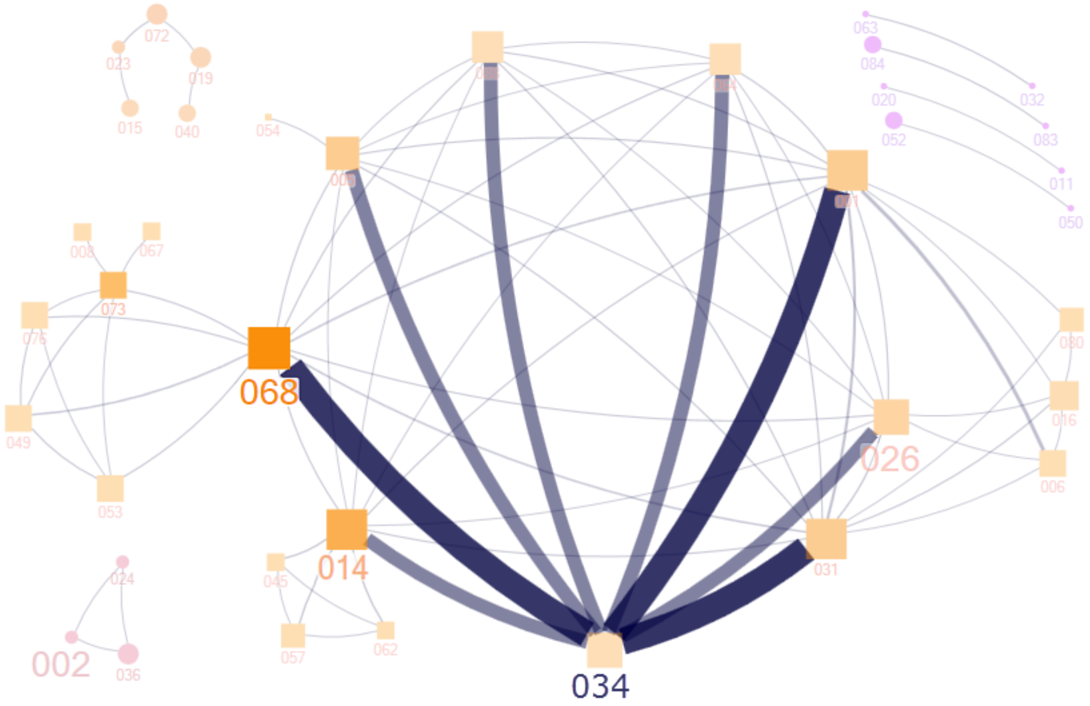


Figure 5.10: Final visualisation of network 6.1

5.2.3 Network 6.1 results

This initial investigation resulted in a clear network graph of the student data, showing all the students, linked to one another by how many files they have that are duplicates or geometrically similar. The final graph shown in figure 5.10 clearly illustrates that some students are highly connected. This is achieved by assigning network metrics to visual properties of the network, which optimise the graph for visual inspection.

At the outset of this analysis edge weight was mapped as width and opacity, shown in figure 5.5, clearly indicating which students have strong connections. The node size and opacity was determined by the node degree and betweenness centrality respectively. Laying out the graph by hand rather than using an automatic layout and changing the colours of the edges allowed clearer inspection of the elements. These visualisations make it very plain which students are most connected or have an important position in the network, information inaccessible prior to this network being made. Once the visualisation was further enhanced and other graph metrics were put to use in node imaging as shown in figure 5.10. The key features relevant to plagiarism detection were clearly evident from a visual inspection.

The node size was determined by the degree of the node, meaning the size of a node indicated the number of unique edges connected to it. A large node shows a student with many other connections, a small node shows a student with few connections and therefore the large nodes are of interest in this investigation. In figure 5.10 the student with the most links to other students is labelled 068 and stands out in the image. The opacity of nodes increases with their betweenness centrality. High opacity indicates which nodes act as 'bridges', showing which student or students connect others together. Nodes which are opaque could indicate a student that is likely to have shared files between other students.

Node colour is determined by its closeness centrality value and indicates the average shortest distance from each node to another. Orange nodes are close to others and pink nodes have a lower closeness centrality value. Despite different levels of opacity the differing colours indicate distinct clusters and show which are closer to others. Separate groups are shown to be less connected, such as nodes 002, 026, 034, through the colour they have been assigned. This use of colour enhances clarity and shows these different groups of students and assists in identifying outliers.

Node shape is determined by the eigenvector centrality, which is a measure of not only the number of connections a node has, but the degree of the vertices it is connected to. Often understood as a measure of the 'influence' a node has in a graph, here it indicates which students are most influential in the network. In figure 5.10 the square nodes have a high eigenvector centrality and the nodes which are circles have a low eigenvector value. This metric clearly indicates which clusters are most influential in the network.

From a combination of colour, shape and opacity it is easy to see which nodes are of little interest in this investigation. The pale pink circles indicate students whose work is not closely linked to anyone else's and who are likely to be involved in plagiarism. Nodes which are small with few connections are also unlikely to represent students who have plagiarised. Their limited connections show they do not have many similar files to other students.

Nodes which are large and opaque indicate students who are more likely candidates of plagiarism. Their strong position within the network shows that they have many files which are geometrically similar or duplicates of their classmates, and this could be further supported by the types of edges they have attached to them.

Both the edge size and opacity are determined by the edge weight, however it is important to note that outliers have been ignored. This was necessary as one node had a self-loop so large that it influence the scale and made other edges appear very small. Large, opaque edges between nodes indicate a large number of similar files have been found between two students, while small, transparent nodes indicate less similarity in the work submitted. Therefore a node which has large, opaque edges travelling to it must show a student with many more similar files to other students, than a node which has a small, transparent edges connecting it.

While a node may appear to have many large, opaque edges travelling to it, this is only indicative of how 'strong' a single link between two students is, not of the total number of similar files a student has. A node may appear to have many 'strong' connections but have a low degree, so still be relatively small in size. The node size must be considered, as this will indicate how many links a node has. Edges therefore indicate how similar one student's work is to that of a classmate, but not how many similar files a student has overall. Visually this can be misguiding, as a node with heavy edges may be perceived to show a problem student. In figure 5.10 it can be seen that student 034 has many heavy edges connecting them to other students. This can be further understood from figure 5.9 where this individual is shown to have a large self-loop. Despite the 'ignore outliers' function being enabled, this edge is still far larger and more opaque than all others, indicating how very high its weight is.

This oversized self-loop indicates that the student had many files that were geometrically similar or duplicated other files they handed in. This indicates one of several things. For example, the student might have handed in two CAD files of a screw that was the same size and dimensions but with different heads, meaning the similarity produced the result. The student may have handed in back-up copies of the file alongside their submission, and similarly, they may simply have duplicates within their own file system, of revisions or drafts and so on. Upon further investigation it was found in this instance that the large self-loop was due to the fact that student 034 had turned in many copies of the same work, with multiple backup folders and revisions of their own parts being included in their submission. This highlights how a network graph alone may not provide clear information about students without further investigation. However, it is notable that in this context a large self-loop is likely to be indicative of this situation and nothing else.

It was entirely reasonable, therefore, to remove these self-loops from the network diagram and concentrate on the other connections between students. However the other noteworthy influence this large self-loop has on the graph is that all other edges travelling to student 034 are very large and opaque. This will once again be the influence of these many similar parts the student's own files contain, for where student 014 may have had originally one similar CAD file to student 034, this will be increase as student 034 has submitted multiple copies or earlier versions of the file, resulting in student 014 having one file similar to several of student 034's. As such a large, heavy edge is shown. These results could be run again, with the backup folders and revisions of students 034's removed from the part collection, however this is unlikely to be of significance at this initial stage. For this reason, contrary to what was first presumed, it is important to not remove self-loops from the initial analysis, else the influence of large self-loops may not be detected and prompt an incorrect interpretation of the data.

From figures 5.9 and 5.10, therefore, we can locate a few key things about this collection of student parts. Student 034 is shown to have the heaviest edges linking them to others and these strongest links are explained by the large number of duplicates within their own files. Student 068 is the most opaque node as they have the highest betweenness centrality and the largest node because they have the highest degree. The highest degree indicates that this student has the most links to other students and also that they are the best bridge in the network.

As this graph does not represent the students' relationship links but the similarity of their work, it is is important to think about the relevance of degree and betweenness centrality and what they really indicate. If the edge weight represents the number of similar files students share, this could be an indication of plagiarism. If degree represents how many similar files a student has to others, a high degree may represent a weak student, who has plagiarised or had help from lots of other students, therefore resulting in a large number of links. If betweenness centrality indicates a bridge student, it is likely they may have passed files between friends. From these initial results it is clear to see that there are a number of graph metrics and measures that may be useful in plagiarism detection.

5.2.4 Network 6.1 conclusions

Combining ShapeSpace's legacy tools to assess the similarity of CAD parts with network theory results is a new technique for analysing collections of CAD models. Using this technique to analyse a collection of student CAD work shows that there are aspects of network theory that are useful for exposing information about the connections between students' work that have never been available before. The visualisations devised are not only novel, but also of use in assessing the collection further.

The software used allows the calculated metrics to be mapped to visual properties of the associated graph. This graph can then be edited into a clear configuration to produce a unique image, allowing the data to be visual interpreted. This produces a novel image where the metrics highlight interesting details.

As there was no academic misconduct found in the class's submissions that were assessed here, it is not possible to determine whether or not plagiarism can be detected from this collection of student work. In fact there is no record of academic misconduct in this or any other historical class within the data archives available to this research. Therefore it was determined that a cheat must be created to investigate whether these network models and the interpretation described above would provide useful information on identifying plagiarised work.

5.3 Network 6.2

Initial Network Model with Created Plagiarism

For clarity the first analysis, which was presented and discussed in section 5.2, is referred to as network 6.1 to indicate it is a first iteration of this process, with the subsequent network being labelled network 6.2, to indicate it has been made with an edited variation of the same data. The notation for the created cheats takes an alphanumeric form, with the network being referred to numerically and the type of plagiarism indicated by assigned a letter. In this section the cheats will be noted as cheat A6.2 and cheat B6.2.

Continuing on from the inspection of network 6.1, network 6.2 was created. Following analysis of network 6.1, the files belonging to student 034 were investigated and many were found to be perfectly innocent drafts (technical drawings). For this reason these files were omitted from the next build of the network, and network 6.2 was created. Figure 5.11 shows network 6.2 after the data has been processed as network 6.1 had been in figure 5.10

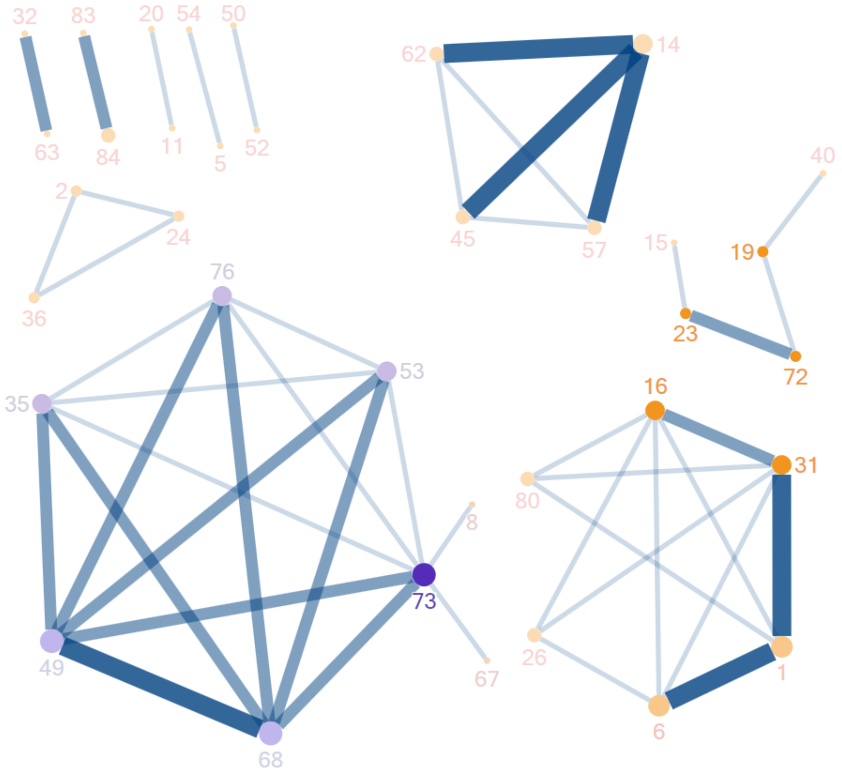


Figure 5.11: Network 6.2

5.3.1 Defining plagiarism

The novel technique for analysing a collection of students' CAD work presented above has shown how the metrics produced from network theory can be useful. The initial discussion of the visual inspection of the graph supported this conclusion. However, while these metrics are useful for uncovering previously unknown information about a class's submissions, no plagiarism was found within this class's results when they were assessed. It is therefore reasonable to assume that no student plagiarised another's work. In order to discern which metrics reliably indicate plagiarism, it was determined to create a cheat from the work the class had submitted, add the cheat to the collection as if this was a genuine student and re-analyse the files, thus creating a new network. It would then be possible to evaluate whether the metrics highlighted the work of the cheat.

It is vital at the start of this investigation to first define what plagiarism is when referring to non-text, 3D submissions. The University of Edinburgh defines plagiarism as

> Plagiarism is the act of copying or including in one's own work, without adequate acknowledgement, intentionally or unintentionally, the work of another or your own previously assessed original work. [186]

This definition can be applied to the students' CAD work in this situation, however it does not fully consider all the practicalities of non-text based work.

Academic integrity in design work is complex because there are no established conventions equivalent to quoting or referencing in written work. Blatant copying of another's ideas is considered improper or unprofessional but there is little clarity about the legitimacy of work that has to some extent been influenced by another person's ideas. The subtle difference between being inspired by another's work and copying material or a concept that is not one's own, is an important issue and is common within non-text based submissions.

In teaching design work, learning by copying is another issue. This is particularly prevalent in the context of CAD education where students may be shown how to create a part via a demonstration and then may copy the model to learn how to use the software tools available to them. It is not considered plagiarism for them to copy the design in this educational context, however it is not clear whether they can call the resultant model their own. The work is their own, however the design is not. It is also unclear whether learning from a fellow student, by copying their work as an example, is to be considered as cheating when learning by copying is the most effective way to be taught.

Something else that is not taken into account by the definition of plagiarism above is the reuse of one's own design. It is often acceptable for a student to reuse a CAD file of their own design in several submissions. For example, a student may make a CAD file of a M3 bolt and hand it in for several different assessments, and this would not be considered plagiarism.

Other considerations and issues within education and plagiarism are discussed in depth in subsection 2.5.4. For the purpose of this work the following definition of plagiarism is proposed

> CAD plagiarism within education is when a student copies or includes in their own work the ideas or actual work of another, intentionally or unintentionally, without adequate acknowledgement.

Another key consideration in this research is that the technique developed and discussed so far does not allow for differing types of duplication to be detected or assessed.

In attempting to detect plagiarism in students' work the duplicated or similar files involved are going to be buried in a large CAD part collection where there are likely to be many close resemblances among files but they may not be engineering, semantic or plagiarised duplicates. It is likely that the duplicates detected will be semantic duplicates as often students will have been designing CAD files for the same brief or project. Also some parts will have the same purpose, no matter where they occur. An example of this would be a screw, that may be geometrically or semantically duplicated.

This investigation is seeking to identify plagiarism that can be categorised as duplication, which may combine geometric and semantic similarity as well as copied design influences or specific ideas. Practically this may look like a student directly copying another student's parts, or taking another's parts and editing them slightly. Therefore the plagiarised parts this investigation is seeking to detect are likely to be copied duplicates, which may have been edited.

ShapeSpace's software reliably assesses the similarity of parts indexed and as such the results from the duplicate analysis it performs can be used as valid data from which to build a network. However further investigation will be necessary to determine whether the duplicates found are truly plagiarism duplicates.

5.3.2 How to create a cheat

The definition in subsection 5.3.1 of plagiarism in CAD provides a framework within which to create a cheat. For the purposes of this investigation only a student who has intentionally plagiarised was modelled. A student who has unintentionally plagiarised by being influenced by friends' work or reusing their own work is more complex to model and the subtleties of unintentional plagiarism were determined to be beyond the scope of this research.

It was decided that a cheat would be created from the class data that had been used to create the initial graph (figure 5.11). This would allow the data to be compared to the original structure and metrics that were produced.

When considering intentional plagiarism it was determined that the main types of cheating would be direct copying of another's work or an attempt to disguise copied work so it would appear original. These two types of cheating could occur independently or be combined by a student who may attempt to take files from one or more class mates who perhaps worked in a group.

Due to this two cheats were modelled, one as a student who had harvested files from several others within the class and one as a student who had worked closely with a friend. These cheats were created by Sophie Broad, who assisted with this work (under the author's guidance) and documented the process extensively in her dissertation [33]. For clarity, these cheats are labelled cheat A6.2 and B6.2, to indicate the type of plagiarism that was modelled and the class they belong to, in this case, the first set of data analysed in this investigation.

The first cheat, cheat A6.2, was created with a folder of parts, some of which were directly copied from other students' files, some of which were copied from other students and then edited and others that were original and newly modelled. This fictional student was not given an assembly, as the part files in the folder did not fit together. It was determined that this scenario would be unlikely if a student was completing the assignment, but more likely if they were attempting to pass by cheating.

It was expected that this cheat would be shown within the network as having many links to the other students, showing who they had copied parts from, illustrated in figure 5.12. The parts that had been copied and edited would also show as links to other students.

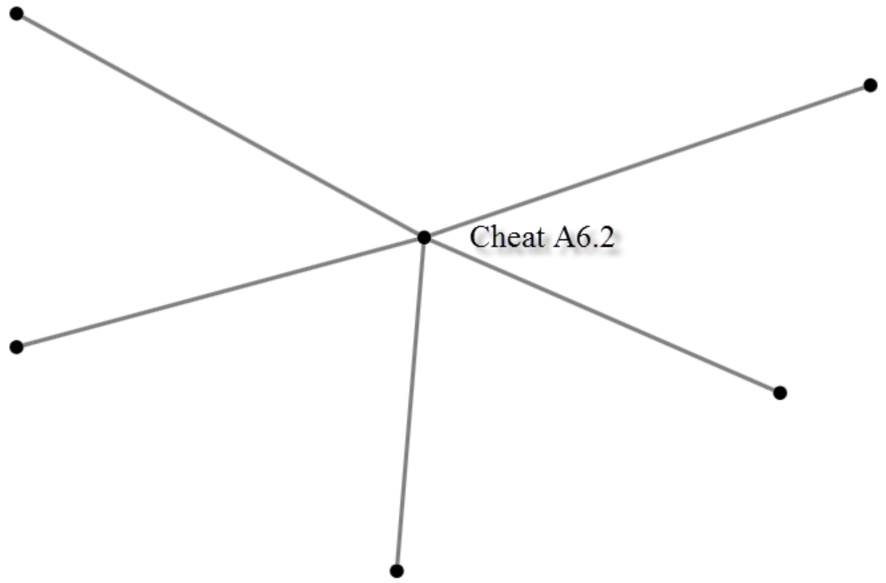


Figure 5.12: Predicted network feature highlighting cheat A6.2

The second cheat, cheat B6.2, was created to model someone who had worked closely with one other student, a close friend perhaps, and directly copied their work or duplicated their design. A folder was made for them containing the same files as one other student, including the assembly file.

The way this student would appear within the network would be as a node with a strong connection (shown by a heavy edge) to one other node on the graph, as illustrated by figure 5.13

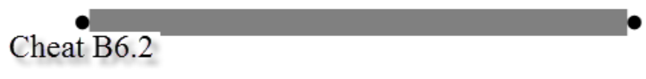


Figure 5.13: Predicted network feature highlighting cheat B6.2

Once the cheats had been created the method presented in subsection 5.2.1 was repeated. The cheats files must be added to the collection and a new network model created, as the software used does not mean simply adding files to the store would allow the connections between those files and the existing ones in the network to be detected.

5.3.3 Plagiarism detection method

Beginning as before with the method described in subsection 5.2.1 the ShapeSpace software was used to analyse a folder that included cheat A6.2 with the original class's work. The method was extended to incorporate the techniques documented in subsection 5.2.2 and so is documented again below for convenience.

The folder containing all the submitted class files as well as the cheat's files is created and analysed with ShapeSpace's similarity software, with each file being indexed before the duplicate analysis is performed. The tab-delimited text file produced by the duplicate analysis is transformed from a list of similar groups to a list of one to one relationships, which is input into NodeXL. The network is visualised and all duplicate edges are counted and merged. The metrics are calculated for the network, the self-loops are not displayed and the metrics are mapped as

- Node degree is mapped to the node size
- The betweenness centrality is mapped to the node opacity
- The closeness centrality is mapped to the node colour
- The eigenvector centrality is mapped to the node shape

- The edge weight is mapped to both the edge size and opacity, with outliers ignored.

This method was preformed once more, where the files made for the second created cheat were included in the folder containing the students' work, and the files belonging to the first cheat were removed.

5.3.4 Results

Figures 5.14 and 5.15 show network 6.2 when it contains cheat A6.2 and B6.2 respectively. These can be compared to figure 5.11 which shows network 6.2 without an added cheat.

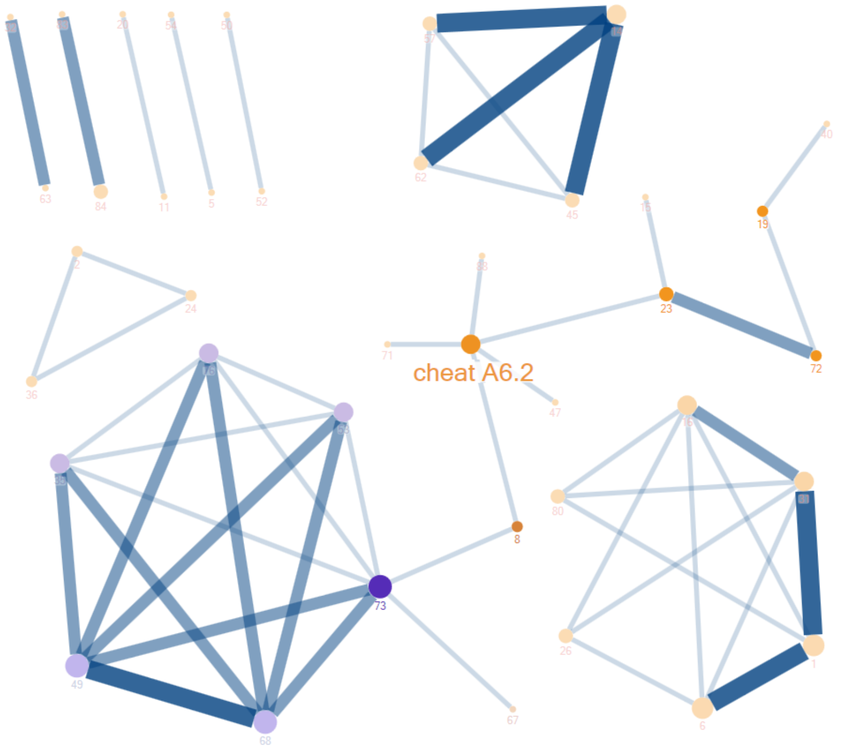


Figure 5.14: Network 6.2 including cheat A6.2

Comparing the three diagrams and their related metrics shows some interesting features and the cheats are straightforward to identify within the network. In figure 5.14 cheat A6.2 has the highest betweenness centrality, shown by the colour of the node and the smallest clustering coefficient. The predicted structure illustrated by figure 5.12 can be seen clearly in figure 5.14, with cheat A6.2 at the centre.

In figure 5.15 cheat B6.2 medium-low degree but a very heavy edge link to student

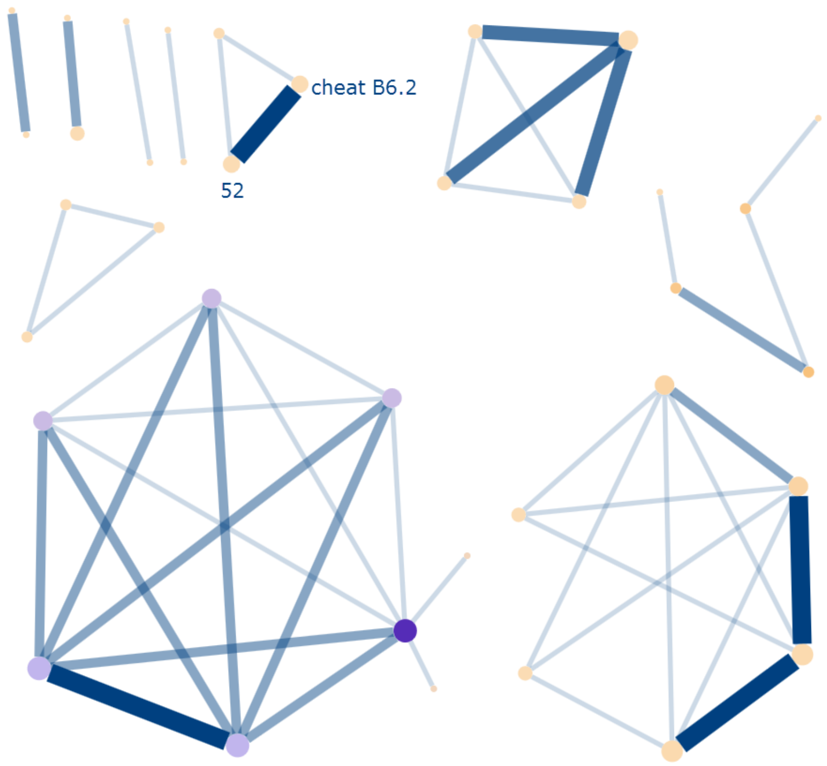


Figure 5.15: Network 6.2 including cheat B6.2

represented by node 52. The predicted feature in figure 5.13 is mirrored by the strong link in figure 5.15, suggesting that the predicted results match the analysis result accurately. Arguably there are other similarly large edges shown in network 6.2 and in a real world setting this may call for those students to be investigated. Notably comparing the metrics of these edges shows that the edge between B6.2 and 52 is larger and more opaque than the others present.

These results indicate that betweenness centrality, clustering coefficient and edge weight values may point to non-original work belonging to students. While these metrics highlight the created cheats put into the network, the features predicted and illustrated by figures 5.12 and 5.13 and clearly shown in the respective networks. These structures were logically predicted due to the parameters the cheats were created within. It is therefore unsurprising that they are clearly identifiable within the networks.

While these results agree with those predicted, it is clear that this structure and the metrics of interest here do not conclusively prove plagiarism has occurred. A clear example of this is shown in figure 5.10 where several students have edges with large edge weights connecting them to student 034 with a large self-loop. It is vital to consider any self-loops in real world applications.

Also these results must be verified to confirm the assumptions used to model the simulated cheats. It was suggested by Broad that the scenario modelled by cheat B6.2 was more realistic than that modelled by cheat A6.2, as a student would likely hand in an assembly, not simply a collection of random parts, making the harvesting modelled by cheat A6.2 less likely as parts obtained from several students would probably not fit well together. The parameters set to model the cheats within were deemed reasonable and realistic, but this must be verified by further simulations.

5.4 Networks 7 and 8

Verifying Initial Results

In order to verify the results found from analysing cheats A6.2 and B6.2 in network 6.2, further cheats were created and inserted into other networks. Networks 7 and 8 were created from historical submission files, associated with the same 3rd year CAD course at the University of Edinburgh. Again for these networks no plagiarism was detected during marking and grading. For both classes the first iteration of the networks were made using the method described in subsection 5.3.3 and showed very little similarity between the students' submissions.

It is notable that these results show lower levels of similarity in the classes used in networks 7 and 8, than the class data used to create networks 6.1 and 6.2. This was determined to be due to the type of assessment set. The assignment set and visualised in network 6.1 and 6.2 was the design of a steering wheel. The assignments related to networks 7 and 8 were the designs for a vernier caliper and a micrometer respectively. It may be argued that submissions for a simple design task would be more similar in construction, given the more limited scope of design possibilities. However it is logical that fewer original files were found to be similar in the case of networks 7 and 8, as the shape similarity software would assess the many geometric aspects of the designs, not their function. This could also indicate a good level of accuracy in the method proposed in subsection 5.2.1, as it finds highly similar parts, rather than vaguely similar parts.

In order for a clear analysis to be made, figures 5.16-5.19 showing networks 7 and 8 have not been edited to remove the self-loops. It was decided that these were important features to include while investigating these data sets, as they are less familiar than network 6.2 shown in figure 5.11, which was assessed in detail in section 5.2. Leaving the self-loops as part of the network diagram would allow them to be considered alongside any other edges that had large weights once the metrics had been mapped to visual properties of the graph.

5.4.1 Cheats A7 and B7

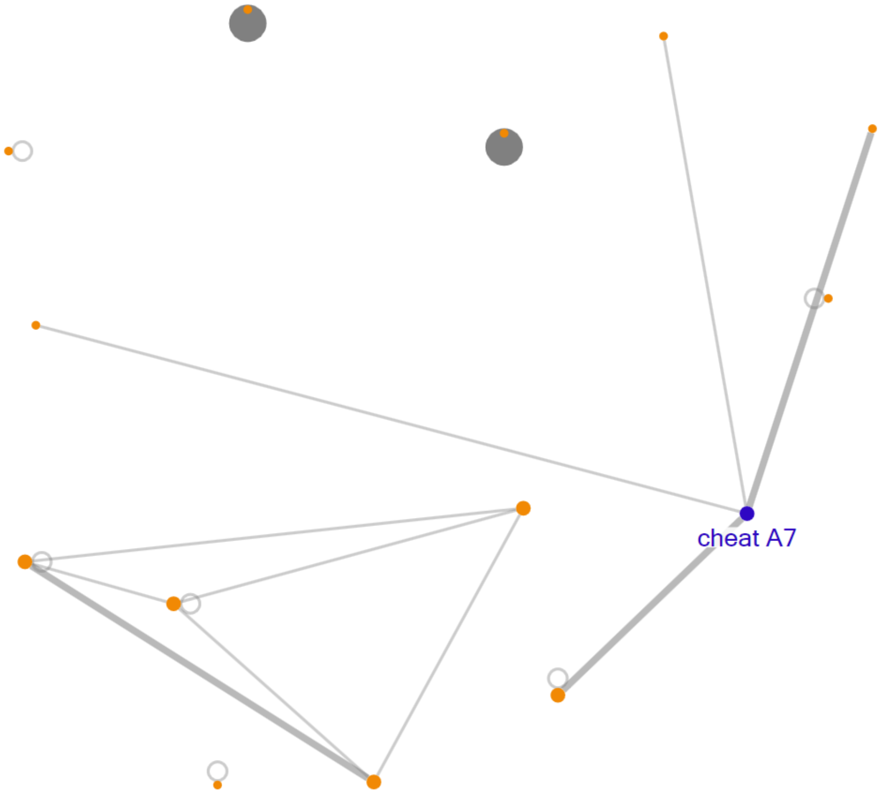


Figure 5.16: Network 7 including cheat A7

For network 7 two cheats were created in the same fashion as cheat A6.2 and B6.2, in order to reliably investigate the metrics that appear to highlight plagiarised work belonging to students. Figure 5.16 shows network 7 including cheat A7. cheat A7 can clearly be identified as a node with a large degree, shown by its large size and a high betweenness centrality, shown by its colour. It has a medium closeness centrality and no other notably high or low metrics. The predicted structure illustrated in figure 5.12 can once again clearly be identified in figure 5.16 and the large degree and high betweenness centrality agree with the results found from cheat A6.2.

Cheat B7 is readily identifiable in figure 5.17, with a large, opaque edge leading between its node and one representing another student. This large opaque edge again indicated that an edge with a large weight matches the predicted structure in figure 5.13 and the way cheat B6.2 was identified in figure 5.15.

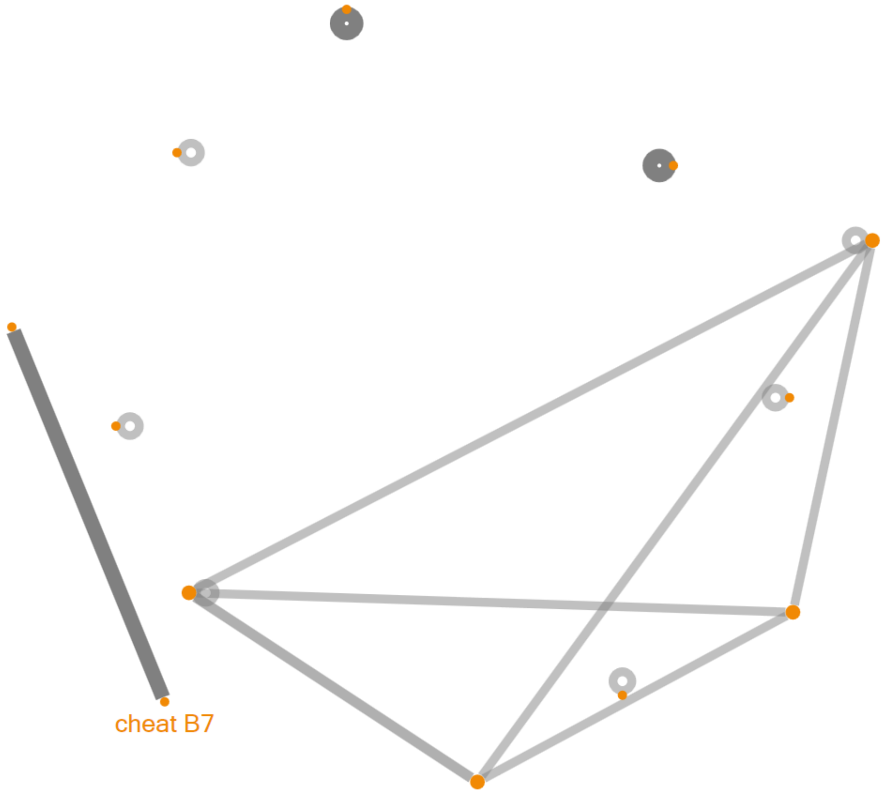


Figure 5.17: Network 7 including cheat B7

5.4.2 Cheats X8, C8 and D8

For network 8 the cheats were created in a more random fashion, instead of within the parameters used previously. They have files that were similar to other students, but highly edited or left as original in a less structured way than cheats A6.2, B6.2, A7 or B7.

The first attempt at creating a different category of plagiarism was to edit a full assembly, and the resulting file was labelled cheat X8. This file was copied and edited to be longer, larger, thinner and have different size screws and holes. Cheat X8 was created to model the situation where a student had managed to steal a final assembly from another student, and edited it so it appeared original. It did not include any copied part files, as the situation modelled was a brief and thoughtless attempt at making the hand-in deadline. However this scenario proved impossible to analyse as the edited assembly file had no part files associated with it, and therefore could not be opened.

For this reason cheat X8 could not be included in the network analysis.

If this plagiarism scenario did occur, it would not be detected by the process developed and presented in subsection 5.3.3 as the file would not be indexed with the other part and assembly files. However this was determined to be a very unlikely situation as it would be nearly impossible for a student to copy only an assembly file and not the dependent parts. Also if this situation did occur a marker would be highly likely to detect it, due to the written section of the assessment being incomplete without the part files.

Instead cheats C8 and D8 were created. Cheat C8 was given a random assortment of other students' work that would not fit together to make a finished assembly, some of which was edited and other files were kept the same. In a similar way to cheats A6.2 and A7 and cheat D8 was generated to include several files from one other students that were chosen and edited at random, with several remaining original, in a similar fashion to B6.2 and B7. These cheats were devised to assess the reliability of the results from cheats A6.2, B6.2, A7 and B7. While they were created in comparable ways they were not invented using the same parameter as the previous 4 invented cheats, and so were expected to return results that are similar but not identical.

Figure 5.18 shows network 8 with cheat C8. Cheat C8 is not as instantly recognisable among the other nodes in the network, with a medium sized node and number of edges travelling to it, indicating a medium degree and a high betweenness centrality. In fact, the node representing cheat C8 has the same degree and betweenness centrality as another node in the network. However the node for cheat C8 also has one very heavy edge travelling to it. While the other blue node does have a heavy edge attached to it, this is also linked to a node with a large self-loop, indicating that the weight of this edge could be due to the large number of similar files that student has submitted.

Due to the network containing fewer nodes than network 6.2, the metrics do not indicate as many interesting features. However once again the cheat is identified by a high value for betweenness centrality and degree. If this was a real world scenario, it is likely a marker would focus first on the cheat's node due to the additional indicator of the large edge, and only then on the other high degree and betweenness centrality node.

Figure 5.19 shows network 8 with cheat D8 included in it. Cheat D8 is much harder to locate within this network, visually or using the numerical metrics. The node representing D8 does not stand out and has no heavy edges leading to it. Overall very little similarity is evident in figure 5.19 with only two small groups of connected

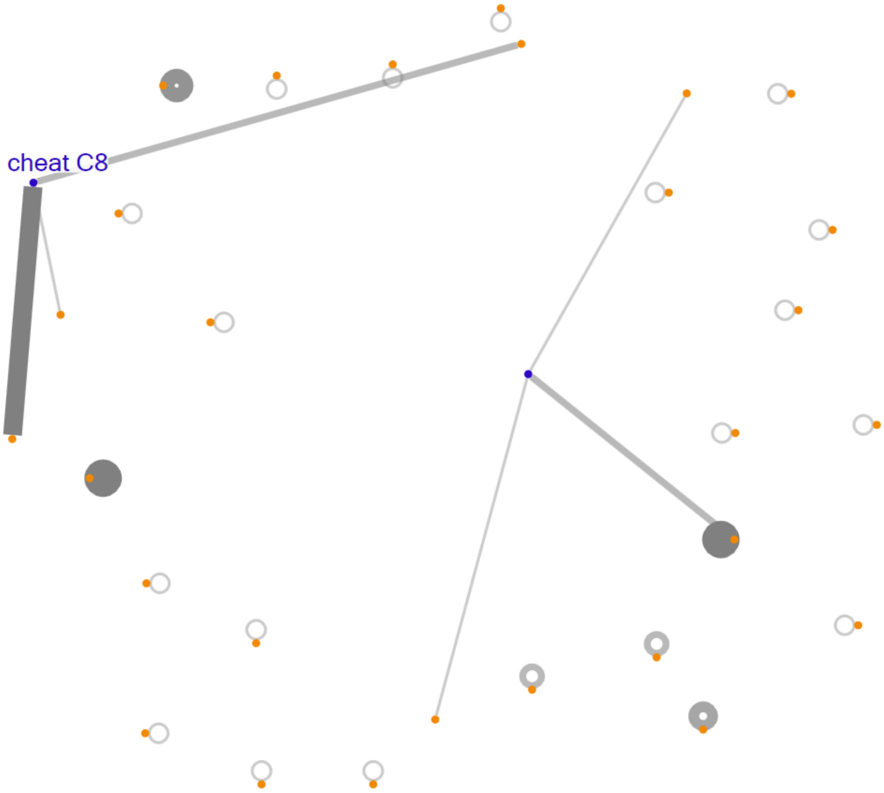


Figure 5.18: Network 8 including cheat C8

students. In this scenario a marker, seeing the network links, would have the much easier task of checking for plagiarism among just 7 students, rather than among all 29.

Cheat C8 is comparable to cheats A6.2 and A7 when visually comparing the networks, however it is not the node with the highest betweenness centrality nor is it clearly the node to investigate, when viewing figure 5.18. The heavy edge connecting it to another node is more comparable to the results of cheats B6.2 and B7. If these results were combined and a scoring system created to indicate who is most likely to have cheated. This would have to be taken into account as it is possible that this type of plagiarism could occur, if a student was finding the work challenging and wished to be seen to submit some form of CAD files; a scoring system could indicate this well. Considering that network 8 had low levels of similarity prior to the cheats being added, this could explain the occurrence of these lower value metrics. However the cheat does have heavier edges leading to it, showing higher similarity with the connected students, than the other blue node in figure 5.18. If this was a real world plagiarism scenario, the cheat would not be the only one highlighted for investigation and human confirmation would still be required.

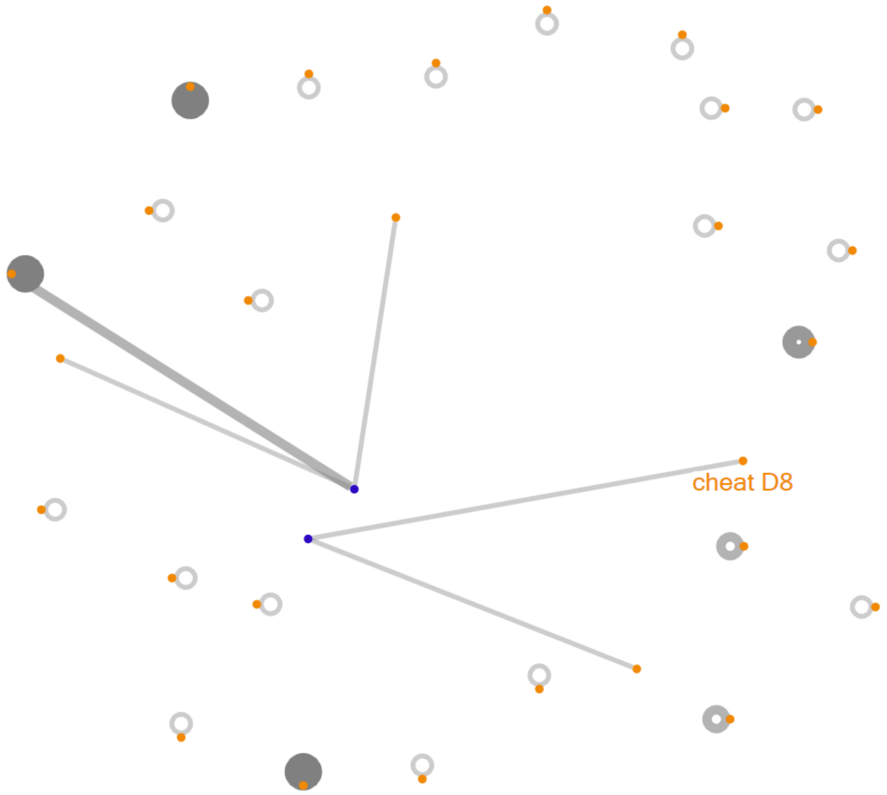


Figure 5.19: Network 8 including cheat D8

5.5 Feature Analysis

Concurrent to this work, Mill has performed an analysis creating networks where CAD models are related by feature use. These networks were created from the Edinburgh Benchmark parts collections and the CAD models were related to the geometric features from which they are made. An example of this work is shown by the bimodal network in figure 5.20 where the red nodes are the CAD parts and the blue nodes are the features.

It was suggested that performing a feature analysis on the students could provide a different way of using network analysis to detect plagiarised work without the need for shape similarity analysis.

This analysis was performed by Sophie Broad [33] and networks were built in the same fashion as employed by Mill from the data on cheats A6.2 and B6.2. The initial network diagrams presented did not clearly highlight either cheat, as the feature analysis contained numerous links between the students' files, as shown in figure 5.21.

Feature graphs with many nodes can be more easily understood when they are split into partial network diagrams. Figure 5.22 shows an example of this created from figure

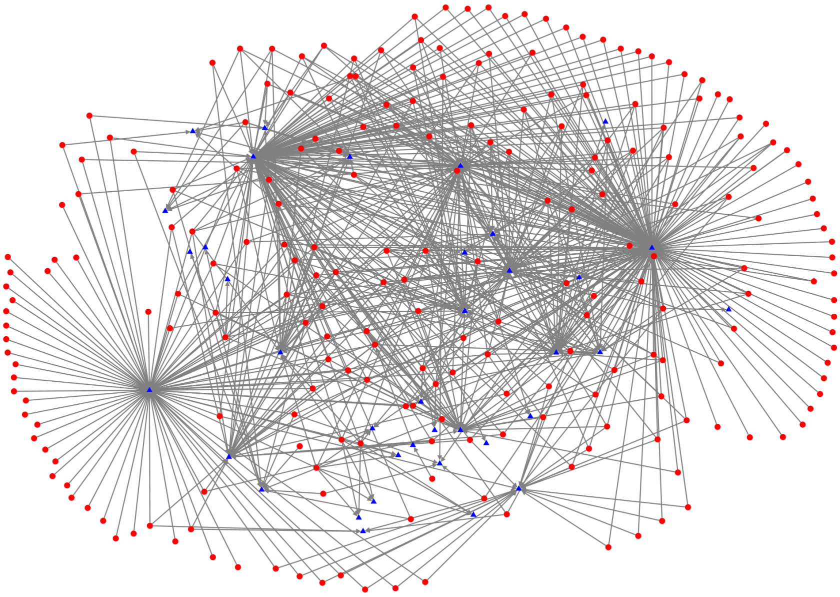


Figure 5.20: Graph of the Edinburgh Benchmark using feature relationships

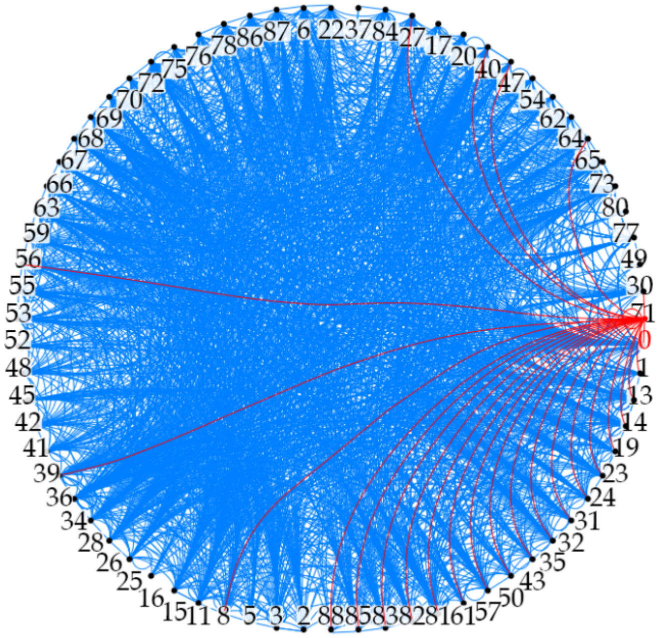
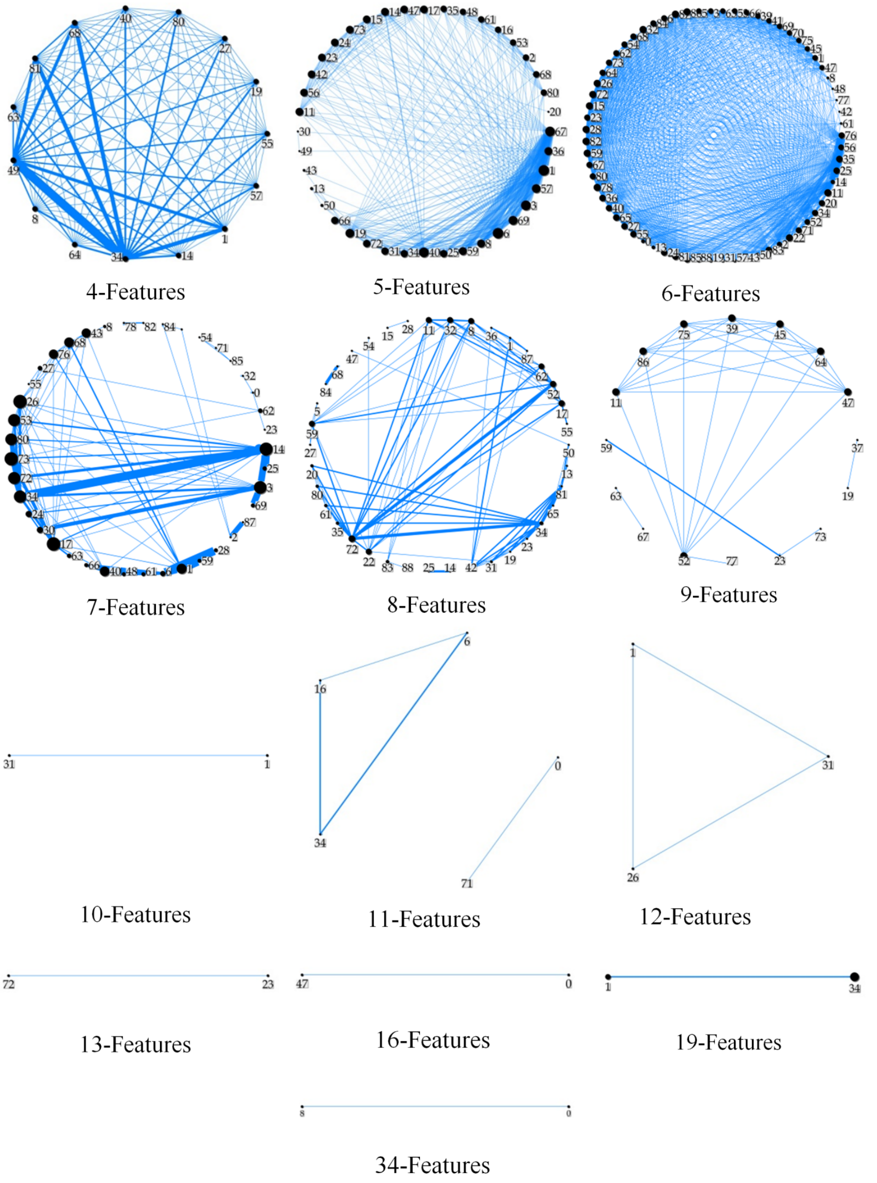


Figure 5.21: Feature analysis network of cheat A6.2

5.21. The network has been split into partial networks that show the students' files as linked based on the number of features they have, indicated by text below each graph. This simplification does not assist with interpreting the data when it is presented in this format, there are still a multitude of connections shown and cheat A6.2 (labelled 0 due to the anonymised data) cannot be clearly located.



4-Features

5-Features

6-Features

7-Features

8-Features

9-Features

10-Features

11-Features

12-Features

13-Features

16-Features

19-Features

34-Features

Figure 5.22: Separated feature analysis network of cheat A6.2

Investigating one student's connections revealed that their clip design file was linked to two students' designs, however files that had been linked in the feature analysis were not geometrically identical. Figure 5.23 is a network diagram that has been enhanced with images to represent the selected student (labelled 049 in the anonymised data) and the way their files were linked to other students'.

This analysis method is less effective in detecting similar work between students when comparing it to the shape similarity network for the same student, shown in figure 5.24 which is the relevant segment of network 6.2.

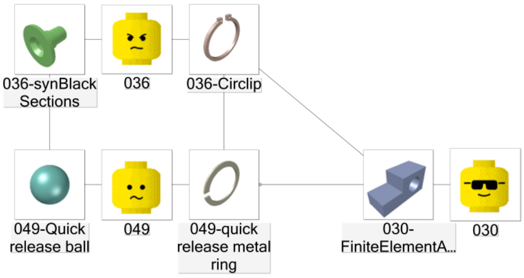


Figure 5.23: Feature analysis network associated with student 049

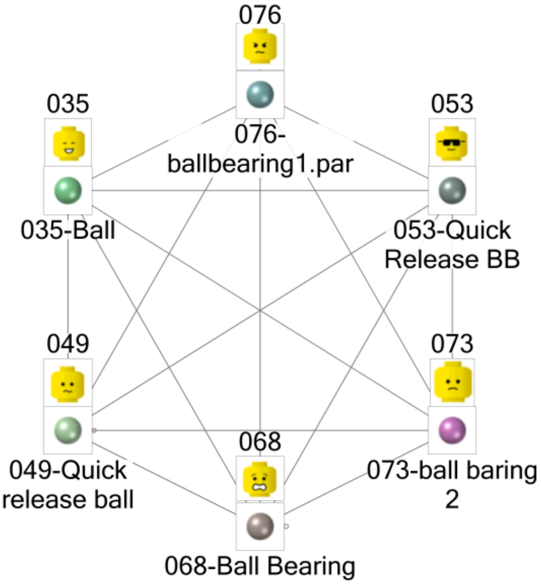


Figure 5.24: Shape similarity analysis network associated with student 049

Figure 5.24 shows six students whose work is not only visually similar, but determined to be geometrically similar within the CAD submissions. It shows that feature analysis of the same data has not found these same parts are linked. The feature analysis was run on the data shown in figure 5.24 and showed only the ball bearings belonging to student 069 and student 076 were created using the same features within the network.

Performing a shape similarity analysis and combining it with a feature analysis could, therefore, give further insight and allow for accurate detection of similar work between students. However on its own feature analysis does not reliably detect similar work or files. Accordingly, feature analysis will not be used in this research to assist in detecting plagiarism.

5.6 Real World Application

It was decided a reasonable concluding step for verifying these results was to use the technique developed in subsection 5.3.3 on a current class. During the examination period in 2015, this approach was used to assess submissions from the third year CAD course and the results were presented to the course organiser and marker and are included in Appendix B. Figure 5.25 is the associated network diagram.

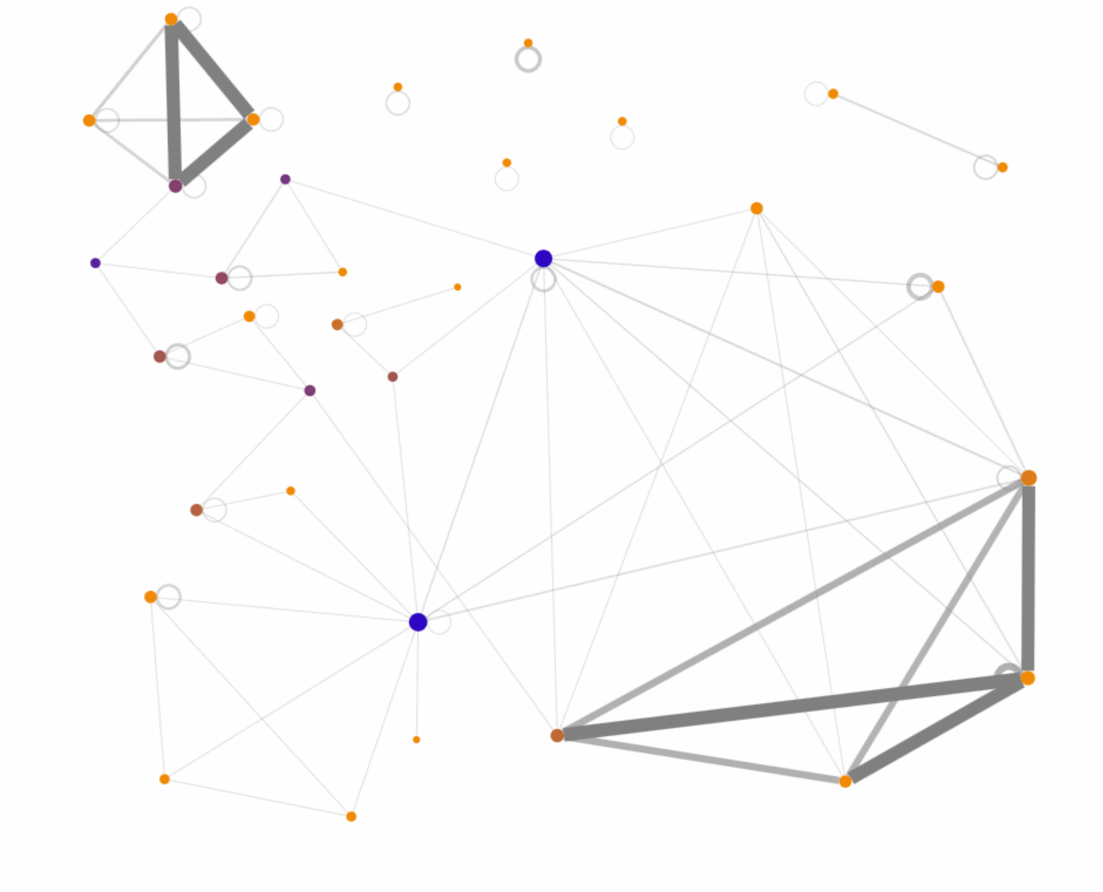


Figure 5.25: Network made to test plagiarism in a current class

As in subsection 5.3.3, the metrics were mapped to the same visual properties. This was done to allow clear analysis and comparison based on previous results. Again self-loops were left in to clarify any particularly large edge weight.

The network analysis highlighted several students of interest. Two students were highlighted as having the highest betweenness centrality and highest degree, shown by the large, blue nodes in the network. These are the same metrics that revealed cheats A6.2 and A7 in their respective networks.

There was one student with a large self-loop and several other large edges connected to it. This is comparable to the results of network 6.1 where a student was found to have handed in a lot of duplicates of their own work.

There also was a small group of students highlighted as having very similar files, shown by a set of heavy edges connecting them. This was the way cheat B6.2 and B7 were revealed, suggesting these students may have worked closely together. These results were discussed and explained in the report that was presented to the marker for consideration (see Appendix B).

In discussing these results with the course organiser, it became clear there were some concerns over one specific student within the class, following an incident in the lab one day during the semester. The results from the network did not highlight that student, however the course organiser's expressed concern was that the design was being copied from a previous year's class, rather than a class mate, and therefore this would not be shown by this analysis. The marker investigated these students and saw there was some clear similarity between the highlighted students. One group in particular had some parts that were geometrically alike, and had similar features, however it was determined this did not amount to plagiarism. Feedback on this method suggested it had aided the marking process, resulting in work being checked for originality in a straightforward manner.

5.7 Discussion

With plagiarism in a CAD context defined, this work developed a method to assess the originality of a class's submissions using shape similarity results to build networks and used the resulting metrics to assess originality.

The initial construction of the network, shown with metrics mapped to visual properties in figure 5.8 revealed the first interesting aspects of the method. Visual inspection, made possible by the enhanced network image, highlighted metrics of interest in comparing the students' work, however the edge weight had to be adjusted, using the 'ignore outliers' function, due to the extremely large self-loop on one student. It was found that this large edge was a result of the student handing in multiple copies of their own work. Therefore the 'ignore outliers' function was used to allow the network metrics to be scaled for easy visualisation.

The self-loops were removed from the graph, as the information these give does not seem to be relevant for the detection of plagiarised work. Removing the self-loops allowed the graph to be simplified, however in further simulation and the real world application of this method, it was found to be useful to retain the self-loops. They then provided vital information about students who had handed in duplicate copies of their own files, not just files that were similar to their classmates. It is clear that if a student has a large self-loop and has similar work to other students, those links are likely to be represented by heavier edges than would otherwise be the case; if a student hands in multiple copies of a part and that part is similar to one another student has submitted, a stronger link will be represented by this method. There is no clear way to improve this, without first processing the files submitted by the students. This would be labour intensive for a marker and would add a lot of time to the marking process, negating the benefit of any time they gain from the similarity analysis. However the real world application presented in section 5.6 shows that the files of the highlighted students can quickly be checked to determine the reason for the heavy edges, and this is preferable to a laborious pre-analysis processing of files.

Creating cheats was an ethical way to assess the effectiveness of this method, as no students had been found to have plagiarised in the classes that provided the original data. The results, illustrated by the network diagrams in figures 5.14, 5.15, 5.16 and 5.17 show that the cheats are easily identifiable. Cheats A6.2 and A7 were created to model a student who had copied files from several other students or sources. The accuracy of ShapeSpace's software, used to perform the similarity analysis, meant that even if the 'stolen' files were edited, they could still be found because of their geometrical similarity to the originals. Cheats B6.2 and B7 were created to model a student who had copied files from only one other student and cheats C8 and D8 were a blend of these two models. An attempt at modelling another type of plagiarism was made with cheat X8, presented in subsection 5.4.2 however this attempt was unsuccessful.

While these successful attempts are not a complete model of plagiarism and the simulated cheats may not encompass every method of plagiarism used by actual students, these scenarios are considered realistic within the CAD education environment. This is supported by the opinion of academics who teach CAD at the University of Edinburgh. Further investigation could include refining the model of plagiarism and a comparison could be made with these results.

The results of these simulations, in every case, show that the invented cheating student can be detected with ease, using the metrics provided by network theory. When these are mapped to visual properties, it is simple to see which students are a concern, identified by their large, bright nodes and the heavy edges connecting them to other students' nodes. The colours of the images were not optimally chosen however, and while the sizes of nodes and edges are easily interpreted, the meaning of the colours are not intuitively obvious so a key is necessary. With this in mind, a revision of the colours and shapes seems appropriate. Many user interfaces make use of a 'traffic light' colour system, where it is understood that red indicates a problem, amber show a warning and green is used for OK. Adopting a red-amber colour scale may allow the diagrams be more easily interpreted. In the same manner triangles are used to warn while circles and squares are used to inform, so it may be suitable to change the mapped metrics to these visual properties. To assess this, the diagram for network 8 including cheat C8 was remade, shown in figure 5.26, using a red-amber colour scheme and different shapes for nodes.

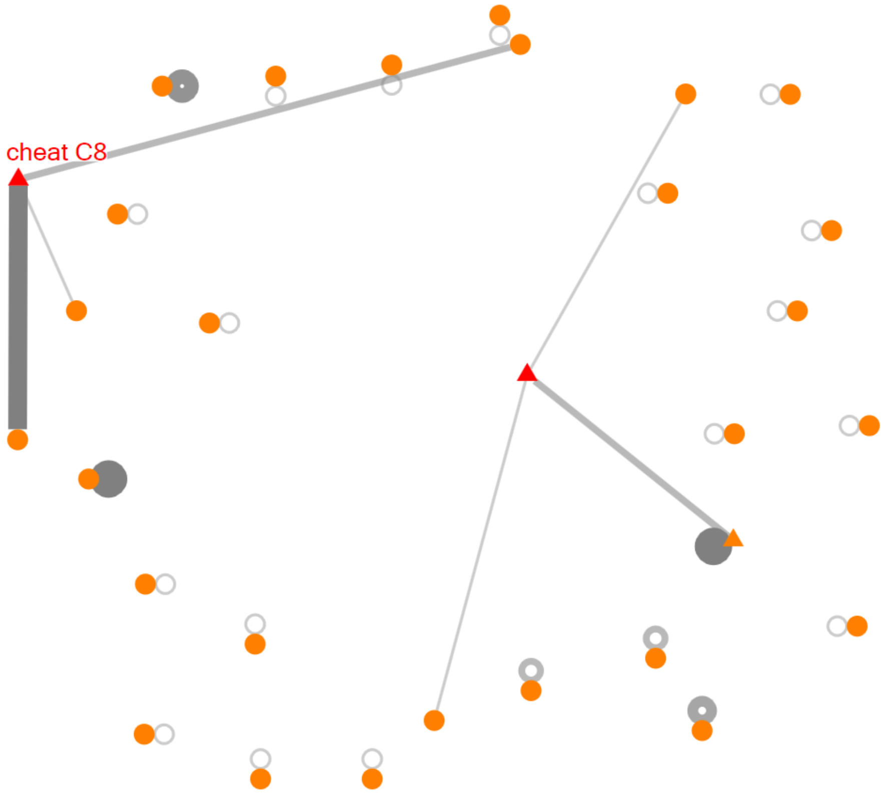


Figure 5.26: Network 8 including cheat C8 with alternative visual properties

The colour of the nodes is determined by betweenness centrality, with red showing a high value and amber for low, and the shape is determined by degree, with a high degree indicated by a triangle and a low degree shown by a circle. The nodes have also been made larger for ease of visualisation. It is much more straightforward to interpret this diagram, with cheat C8 shown by a red triangle. There are two other triangle nodes in the diagram, one of which is also red. These visual changes make it simple to interpret the results, without the need for a key.

The results from the simulations suggest a reliable indicator of plagiarism. In the real world assessment, presented in section 5.6 the developed method was applied to a current class. The results highlighted two students as potentially having unoriginal work, as well as a group of students. These results were written into a report and given to the marker for this course, who said the report made it easy and simple to look at students' work, who maybe had cheated. No plagiarism was found to have been committed, as the students who were highlighted were investigated and, while they did have similar parts, it was not a concerning amount. The feedback received on this 'originality report' stated it made the job of marking more straightforward, and there was no concern over any unoriginal work being assessed. In this respect these results can be seen to have aided academic integrity.

In this chapter the accuracy of the shape similarity analysis has not been considered. As the program provided by ShapeSpace is robust and used in industry, it is justifiable to accept its results. However there were some simulations where a portion of the CAD files were not successfully parsed. This is outwith the scope of this work and it is reasonable to assume, due to the nature of network scaling, that the metrics found to highlight unoriginal work are accurate. Also this method presented in subsection 5.3.3 has not been used to assess a network where two invented cheats were present and the results for finding cheat C8 are not as conclusive as those for cheats A6.2 and A7. While C8 lacks the clear indicators the earlier results show, the cheat is still identifiable by the heavy edge travelling to it, conspicuous within the network diagram (figure 5.18). This result indicates that even if the cheat is not identifiable by all or several of the strongest indicative metrics, within their class they are clearly discernible. It is reasonable to assume this would be the case where a class contained multiple cheats.

The metrics identified as effective in locating plagiarism are shown to be robust through the multiple simulations and the real world exercise. Feature analysis, while able to enhance the results delivered by the network analysis, is less robust since it cannot discriminate the presence of plagiarism from other factors. It would be possible to combine feature analysis with network metrics to enhance the detection of unoriginal work, but that may be unnecessary. The real world application confirmed that the network analysis on its own supplied sufficient information to aid a marker. Incorporating feature analysis would involve more processing time and could further assist a marker in determining unoriginal work. It is worth noting that computer analysis is widely considered unable to provide a sufficiently accurate or reliable measure of plagiarism, so human interpretation is needed to confirm the presence of unoriginal work. This is seen in the real world application where the marker was able to use the results, but the computer analysis was inconclusive. This corresponds with the opinion held by other practitioners, seeking to uphold academic integrity within non-text based subjects [51].

While this work has found an effective way to detect unoriginal work, the literature discussed in subsection 2.5.4 would suggest the issue of plagiarism within education will not be solved by detection methods. If the method presented in this research were to be developed into off-the-shelf software, there could be several advantages to CAD education. The developed software could be given a user-friendly interface, and be made available to educators and students. This would allow lecturers and other educators to check students' work and students could have similar access to it, allowing them to check their own work before submission. There are other associated problems with this however, including students spending time developing ways to trick the software, which has been reported with turnitin, or the concern that only conscientious students would use the checking function while those who were likely to plagiarise would simply ignore.

5.7.1 Critique of recent similar work

Other work focussing on CAD plagiarism was performed by Houjou (2013), who introduced an evaluation method and claimed it prevented plagiarism, by preventing the copying of another's work [85]. Their method was to provide students who were falling behind with sample solutions and thus negate the need for students to plagiarise. Houjou's method however, may not be applicable across all CAD education and while preventative, does not solve the problem of detecting plagiarism. The work presented here may be more widely applicable in CAD education.

Krüger and Wartzack (2014) briefly discuss a method for detecting plagiarism in 3D CAD submissions, while attempting to identify those solutions which are incorrect, using an MD5 checksum and other meta-data. They state that if the MD5 checksum or 'fingerprint' of two CAD files is the same, the parts are identical and in this way copied parts can be detected. However the MD5 checksum changes for a file when that file is saved again. Renaming the file does not change the MD5 checksum, but saving it and making only basic changes would change the fingerprint drastically [110].

Krüger and Wartzack go on to use simple geometric shape comparison to analyse the submitted STEP files. Their aim is to identify incorrect solutions to the set assignment. This method compares each submitted CAD model individually against a reference solution using the geometric characteristics of size of the bounding box, surface, volume and centre of gravity. They analyse 657 submitted files but do not report the time this computation took. From here they split the results into different categories which then require manual review. They do not use these geometric measures to compare the similarity of parts, only to compare them to a correct solution. They defend not using a more in-depth geometric analysis as the aim of their analysis is to enable them to provide feedback to students whose solutions are incorrect.

Given that plagiarism is not confined to students who submit copied files from classmates, their method is neither reliable nor robust. It is questionable whether students aiming to cheat in this way would make no changes at all to the files, which their method would not be capable of detecting. Furthermore, Krüger and Wartzack do not attempt to discuss plagiarism issues caused by parts that have been edited or changed.

Krüger and Wartzack assert that a student with basic IT knowledge could evade their MD5 checksum-based approach and more a robust criterion is needed to detect plagiarism accurately, especially as they were surprised by the high levels of plagiarism they detected. In 657 file submissions, they found 69 cases of plagiarism on the basis of their MD5 criterion and claim that much of this was confirmed by student's admitting they had submitted a copy of a fellow student's work. There are few results and little other work on this precise topic to suggest whether their findings of high levels of plagiarism are typical, especially as the results reported in this research do not concur. However the courses being assessed in these two cases are comparable only because they require 3D CAD submissions. The aims and focus of the two CAD courses are observably different and this may explain why differing levels of plagiarism were found. Other factors that could contribute to this discrepancy may include the systems used, the cohort year, the educational institution, the teaching and other staff involved with the course and the availability of similar solutions to be found online.

It seems reasonable to consider the method reported in subsection 5.2.1 is more reliable than the method used by Krüger and Wartzack, as they are only able to detect precisely identical files that are submitted by different students and the method presented in subsection 5.2.1 allows a broader comparison of the submitted files. Furthermore, the four geometric measures used by Krüger and Wartzack are included within the thirty measures ShapeSpace's software assesses. Also the comparison used to assess this collection of measures has been proved reliable by ShapeSpace's experience using this software in industry, while the comparison used by Krüger and Wartzack has not been validated. The simplicity of their method, however, could be seen as an advantage. Simplifying the comparison method could be beneficial to the process of detecting similar work, but may compromise the accuracy of the results. It is therefore reasonable to continue using the proven software supplied by ShapeSpace.

5.8 Conclusions

The aim of the research in this chapter was to explore the uses of network theory when applied to collections of 3D CAD models by

> Developing the uses of networks in an educational setting, by investigating a class's design submissions, focussing on the uses of networks and shape similarity techniques to assess the similarity of students' work, and thus develop a way to identify plagiarism from a class's submissions (From aims and objectives in section 1.3)

and that has been achieved by building networks based on CAD data from classes at the University of Edinburgh. With plagiarism in a CAD context defined, this work developed a method to assess the originality of a class's submissions using shape similarity results to build networks and the resulting metrics to assess originality. Useful network metrics for detecting plagiarised work have been identified.

Building on chapter 4 the work recorded here developed the uses of networks, built from the results of shape similarity assessment of students' CAD coursework submissions. This is an entirely novel concept and the method detailed is also a unique way to assess coursework from a non-text based subject. The data used to perform these simulations was obtained from past CAD classes at the University of Edinburgh. Cheats were created from class data, modelled effectively on realistic plagiarism situations. Networks were then created from the actual student submissions with the cheats added. It was found that the network metrics were able to pin point the cheats within the network. This was simulated 6 times and all results agreed that the metrics that highlight plagiarism are:

- Large betweenness centrality
- Low clustering coefficient
- The presence of heavy edges connecting students

This was tested again with real world data. The same method was used to assess a recent CAD class at the University of Edinburgh and the three metrics listed above were used to write a report on likely plagiarism. This was found to be helpful, by the marker, and the report meant the students of concern could be quickly assessed to see if there was any actual plagiarism. This real world trial confirmed that the three metrics were helpful in assessing plagiarism in a real class's submissions.

The visualisation of student CAD data in this way is also an entirely novel concept, uncovering information that was previously inaccessible to educators. Mapping the metrics to visual properties of the network allowed the useful metrics to be readily identified.

Feature analysis, performed in cooperation with Broad and Mill, was found to be ineffective at separating out cases of plagiarism. The network built of CAD parts related by feature use, where the nodes represent the CAD models and the edges show which CAD models use the same features, is another intriguing concept, but does not add to the fundamental functionality of this test. As human interpretation is needed to confirm plagiarism, it was deemed unnecessary to include additional feature analysis.

While this method has achieved its aim and demonstrated that it is possible to detect plagiarism, it does not address the issue of plagiarism as an educational or moral concern. A discussion of this is presented in subsection 2.5.4 and it is unclear whether these result will help address the issue of cheating within the educational system. They will, however, aid educators in marking with academic integrity.

5.9 Further Recommendations

Research could continue in this area to more adequately define plagiarism in 3D CAD models. This could then be improved upon, by defining plagiarism in non-text work. This would allow standard practice and codes of conduct to be put in place. Beginning with a definition of plagiarism in a non-text context would provide a platform for a unified approach across higher education. The Oxford Dictionary defines plagiarism as

“The practice of taking someone else’s work or ideas and passing them off as one’s own” [56] and Culwin and Lancaster define student plagiarism as “Plagiarism with the intent of gaining academic credit” [51]

This work proposes the definition for plagiarism in this 3D mechanical CAD context to be

> CAD plagiarism within education is when a student copies or includes in their own work the ideas or actual work of another, intentionally or unintentionally, without adequate acknowledgement.

However, this definition is specific to engineering design only. Improving it to include other forms of non-text based work would allow a unified approach aimed at combating plagiarism across design and image based subjects to be formed, enabling universities and other educational institutions to tackle the issue of plagiarism in CAD education, as well as other non-text based subjects.

To improve upon the work above, the method could be made into off-the-shelf software that would present results to a user. This would allow lecturers and other educators, as well as students easy access to these developments. The software could show the user the resulting network diagram, however this would require a level of understanding on the user’s part, for them to be able to interpret the graph correctly. It would be much simpler if the results were simplified to show a percentage of how similar the student’s work was, in the same fashion as the turnitin method, instead of an image that would require prior knowledge or training to interpret correctly. This would mean the detection method developed here could be easily used and understood by educators.

This percentage could be calculated from the metrics, and combining them correctly could give a reliable percentage of similarity between students’ work. There is no obvious way this could be achieved, however, as the students’ parts are already compared for similarity, so this would need careful consideration. While some of the metrics may indicate a student who has plagiarised, they do not conclusively prove it. For this further investigation drawing on human expertise is necessary. In short, a similarity value must be approached with caution. This work, while effective in detecting unoriginal work, has not proved that plagiarism can be conclusively determined, only that it can be indicated using this method. As such further work could investigate whether such a measure is obtainable or accurate, and whether it would improve the indication of plagiarism network theory allows in this context.

Another key element for consideration would be the inclusion of self-loops and adjusted edge weights, as they reveal important information about the class's submissions. If the network diagram was removed from the results presented to the user, it would still be important to find a way to include the information that self-loops reveal, as discussed in the treatment of the real world results in section 5.7.

If this method were incorporated into off-the-shelf software for educational institutions it would allow lecturers and other educators to assess the students' work effectively. The software could also be expanded to include more CAD models, such as those readily available online at GrabCAD and previous years' work, to give a more robust plagiarism analysis. This could also be an effective tool in advancing the use of references in 3D work, where students could learn how to reference their design influences.

Chapter 6

Multimodal Networks

Investigation of networks built from CAD data in an industrial setting: Analysis of assembly and PLM level data by multimodal networks

In the previous chapter CAD in an educational setting was investigated and it was found that network theory provided an original methodology for detecting plagiarised work. Chapter 5 focused on unimodal network analysis, while the discussion in section 4.2 and throughout chapter 4 highlighted the complexity of analysing multimodal networks. Simultaneously the value of CAD data, particularly within an industrial setting, has been presented in section 2.5.

In this chapter an investigation into CAD data from an industrial setting is presented, with a focus on exploring the structure of company CAD in assemblies and PLM systems. Graph database methods, using TinkerPop, and network science methods were used to analyse anonymised, unedited data from two real world companies, company A and company B. Network 5 from chapter 4 was created using real world data from company A and was considered when the work introduced in this chapter was undertaken. Graph database methods were found to be effective in analysis of CAD data. Multimodal networks are considerably more complex to analyse than unimodal networks, and as such provide a novel way to model real world manufacturing and BOM data.

6.1 Introduction

CAD data, particularly from manufacturing companies, is often considered the most valuable 3D data [47, 85, 110, 204]. The complexity of this data and the systems in which it is contained have become the subject of much work, but few improvements have been achieved (subsection 2.5.3). The structure of CAD data was explored in section 4.3 where assembly structure, as well as the structure of data within an industrial company was briefly discussed.

Multimodal networks, also discussed in chapter 4, may provide tools for modelling this complex company data. The nature of engineering design and manufacturing means that all designs are naturally linked to orders, purchases, customers and sources, from individual components to large assemblies. It is these links that make multimodal networks an ideal vehicle for this data modelling task. For assessing real world company data network theory was applied directly or used as the underlying capability by which investigation was performed throughout the work presented in this chapter.

A multimodal network is where the nodes represent different types and categories of 'things', while the edges represent different types of relationship between these nodes. For a multimodal network any associated metrics must be carefully considered, as the nodes representing non-uniform data may be easily misinterpreted. If a multimodal network is connected as one element, it is possible to measure the diameter and density of a network; specific node metrics, however, may not be so straightforward. An example would be correct interpretation of the clustering coefficient of a node, traditionally thought of as a measure of how connected a node's neighbours are to one another. In this case it would be imperative to consider the influence of different types of node on this metric. This would be unnecessary in a unimodal network as the nodes all represent the same type of data.

Previously this thesis has viewed networks simply, supposing that CAD files can be assumed to be of the same 'type' to allow straightforward interpretations of networks 1, 2, 3.1 and 5. In this chapter however this assumption will not be observed and instead networks that contain individual components, sub-assemblies and assemblies linked by 'contains' relationship edges will be considered bimodal, while those with additional node types and edge relationships will be considered multimodal. As this chapter builds on chapter 4 it is speculated that correctly identifying bimodal and multimodal networks will shed light on the data they model, which would not be evident if the simpler view of networks was continued.

Working with ShapeSpace, it was possible to perform research where network theory was applied to current collections of company CAD and PLM data. This allowed the methodology used in chapters 4 and 5 to be explored alongside real-time industrial analysis. The research presented in this chapter was motivated not only by the findings of previous work, but also actual industrial issues and needs. Research was conducted on real world collections of parts which were unedited and, in accordance with industry requirement, explored the uses of network theory in assisting CAD collection assessment.

6.1.1 Aims of multimodal network exploration

For the initial investigation it was determined that a network should be built from the real world data of a mechanical engineering manufacturer. These CAD files and other associated data were accessed anonymously through collaboration with ShapeSpace. It was determined that summery networks should be built to allow an overview of the company's data structure.

The aims of this chapter's investigation are to:

- Assess the usefulness of modelling real world industrial CAD data.
- Assess the use of network theory in describing a collection of CAD models and associated data.
- Assess the use of network theory in assisting with data analysis needed in industry.
- Assess the use of network theory in advancing shape search and design reuse methods.

6.2 Industrial Data Structure

The most notable difference between CAD in industry and educational situations is the structure of the data. While CAD collections in both areas will have similar assembly structures, where top level assemblies can be linked to the sub-assemblies and individual components they are composed of by 'contains' relationships, and can be assessed for shape similarity, CAD in industry will have additional data directly related to it that is absent in education. Also, it should be considered that there is no single way things are done within engineering manufacturers, and often when discussing manufacturing a scenario is characterised, providing context for the analysis. Network analysis provides techniques that can be widely applied and therefore the CAD collections modelled do not need to be characterised by their context.

As CAD collections in mechanical engineering industry belong to companies who design and manufacture products, the files related to the products will be linked to or contain metadata about materials and specific manufacturing information that would not be present in educational CAD files. This information may be included as metadata in the actual CAD files or may be contained in a Bill of Materials (BOM) associated with a final product. A BOM is not just a file containing assembly information, but also details about materials, labour, sizes and associated quantities.

In addition to this directly related data, CAD collections within a company will be linked to the orders placed for them, by top level assemblies or final products, as well as to the customers placing the orders. CAD files will be linked to the design engineers who created them, and there will be a chain of responsibility linking the manager or engineer responsible to the final products. In some cases these files will be linked to external purchases, where parts are bought in, but still modelled to represent a complete final assembly. Figure 6.1 is an example of this data structure, with all these predicted elements of data included.

The expected network in figure 6.1 is a directed multimodal network that has been arranged by hand for clarity, where node size is determined by the corresponding in-degree value and node colour is determined by closeness centrality, with red indicating a low value and blue indicating a high value. These metrics were chosen to identify influential and important nodes within the network. In the centre of the figure a large blue node labelled ‘CAD collection’ can be seen, with a self-loop edge modelling a ‘contains’ relationship. The size of this degree shows it has the highest in-degree in the network, as it is predicted that the collection within the data structure will be the most heavily influenced by other nodes, while its blue colour represents its central position within the network. While the self-loop edge may be initially confusing, this node and relationship can be understood as a simplification or bird’s eye view of networks such as network 5, where a large collection is modelled by edges representing ‘contains’ relationships linking nodes representing assemblies, sub-assemblies and components. In order to provide a clear overview of the data structure this has been simplified and is represented by one node.

In figure 6.1 the most influential nodes are the customers and responsible engineers,

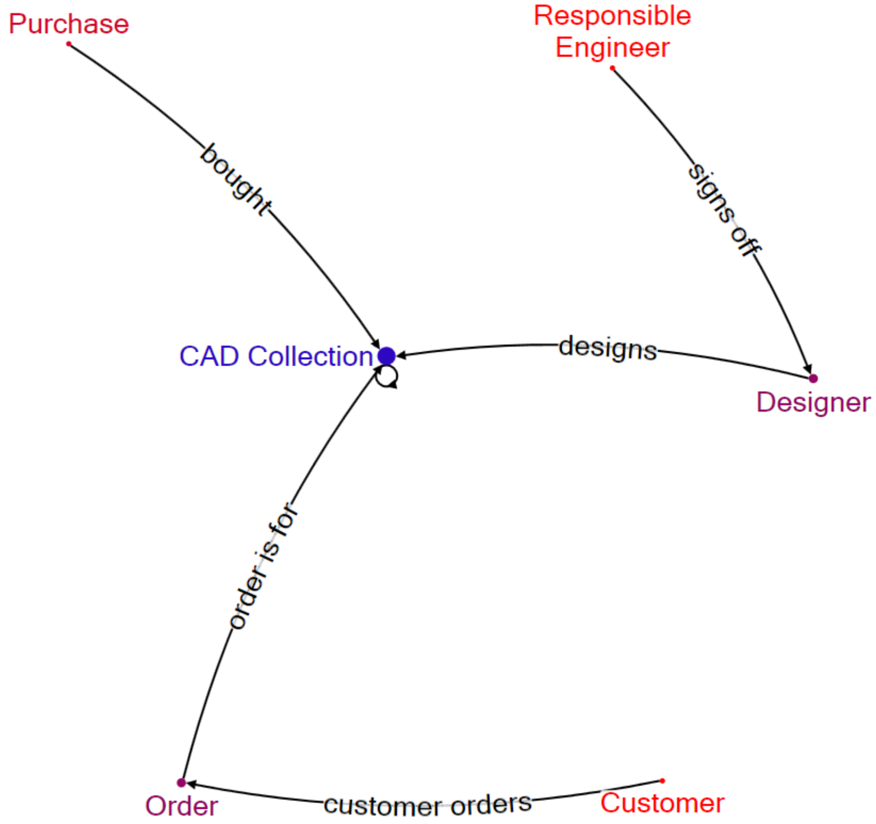


Figure 6.1: Expected overview network structure of industrial CAD Data

as it is predicted that these two types will affect other elements of the data structure, namely the orders and designers respectively. Also included in the network is the influence of purchases. The edges of the predicted network are clearly labelled, and represent the flow of information or influence within the data structure. It is expected that as industrial CAD collections are investigated they will exhibit similar structures to that shown in figure 6.1 and that the BOM will exhibit a similar structure to that of network 1 (section 4.2, figure 4.2).

6.2.1 Bill of Material data structure

Before investigating and discussing industrial data, the basic structure of an industrial BOM should be clearly understood. In manufacturing industry products are most commonly documented and communicated using BOMs, which are lists of product structure, including all details on sub-assemblies, components, raw materials and quantities. Within industry BOMs differ and even within a company a design BOM is different to a manufacturing BOM; a manufacturing BOM may include lithium grease, used to make a thread, but a design BOM would not include this entry. Therefore, figure 6.2 shows a network representing a standard BOM, accepted as a model of an industrial BOM that includes the most common elements.

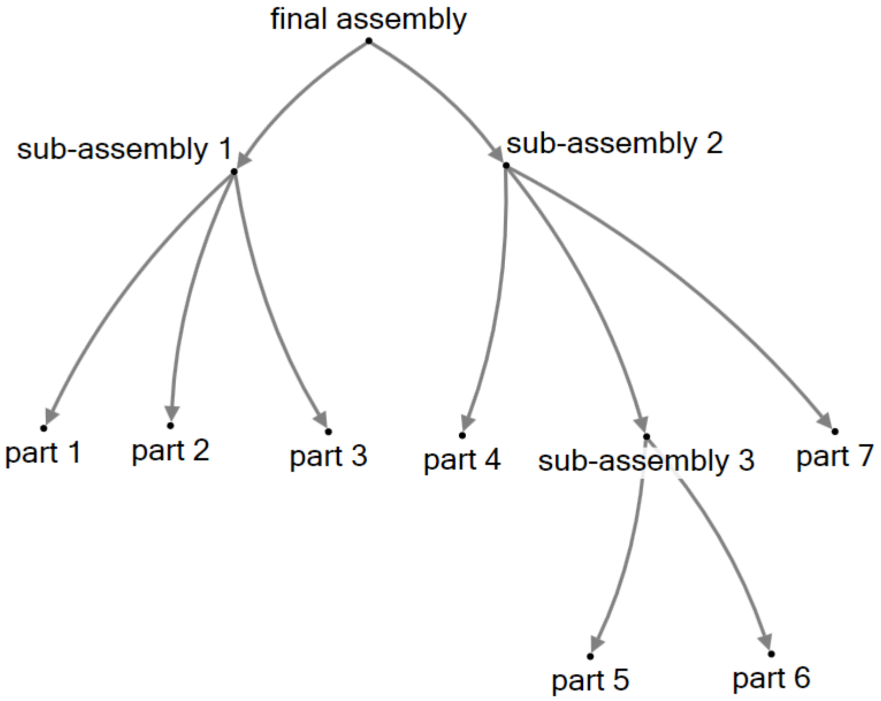


Figure 6.2: Graphic representation of a standard BOM data structure

Figure 6.2 is a directed bimodal graph, arranged by hand, where nodes represent assemblies, sub-assemblies and components and the edges linking them represent a 'contains' relationship. The structure could have also been generated using the Sugiyama forced layout. At the top of the graph the highest node shows the final top-assembly while the lowest nodes represent the single CAD parts. The network clearly depicts how the final assembly is created, with a composite of the nodes below it. Assembly instruction information may be better represented by reversing the direction of this graph, where the meaning of the links could be maintained and noted as 'assembles' or 'contained within'. This type of image is sometimes called a tree structure, where the single components, represented by the nodes on the lowest level, are referred to as leaves.

It is possible to gain an impression of the complexity of the final assembly in figure 6.2 - it is conveyed by the depth of the graph (3 levels represented here) and the breadth (there are several further small tree structures underneath the main assembly). Comparing this basic representation to networks 1 and 2, presented in chapter 4 and shown in figure 4.2 and 4.4 it can be seen how a network diagram can communicate how complex an assembly or BOM is. Modelling this data structure, along with other associated metrics, may allow a comparison of different assembly structures within the same company.

This network style echoes a traditional assembly instruction, and the common tree structure, where engineers use similar methods to communicate how to assemble a final product. While network theory provides a novel basis for presenting this data, the final results and visualisations are comparable to traditional and accepted methods. As such network theory provides and underpins novel methods but results in similar visualisation and data communication. Figure 6.2 is a tree representation of BOM and assembly data structure, produced in a novel way, and is relevant as a simple first model as this research continues to explore industrial CAD data.

As the network shown in figure 6.2 is similar to networks 1 and 2, from chapter 4 (figures 4.2 and 4.4) it is evident that a single assembly, several product structures, or a single BOM, are much simpler than a whole CAD collection. This is clearly illustrated when comparing networks 1 and 5 (figures 4.2 and 4.7) where it can be said that network 1 is an example of the many assembly structures which are contained within network 5. Similarly when assessing large CAD collection in industry, it is notable that they are made up of many BOMs, often hundreds, and this greatly affects the data structure.

6.3 Company A Data Structure Investigation: Network 9.1

Company A is a large engineering design company specialising in power distribution switchgear, with a focus on utility, industrial and commercial applications. Company A engaged ShapeSpace in a large consolidation project, with a particular focus on one key product, for the purposes of this analysis referred to as product A. Product A was so large that it accounted for around 40% of company A's revenue. When company A approached ShapeSpace the data for project A was split between several BOMs; a list of bought-in parts, a list of in-house produced parts and a list of new BOMs produced over recent years. These lists were not definitive and had components which were incorrectly sorted and duplicated between them. ShapeSpace consolidated these lists into one collection and, using their product intelligence services, improved the quality of the company's data collection.

This is a typical problem within engineering manufacturing companies, where legacy products are designed and fabricated alongside new products for existing and new customers. It was known that the design engineers from company A did not consistently store data on materials in one place, further complicating project A. Some of this data was stored as text on technical drawings and some was stored in a new list of material and finish attributes. The material data included as text on drawing files was inaccessible and often read 'see BOM', instead of being assigned as an attribute of the component, as it was considered common knowledge for design and manufacturing personnel working on project A. As part of the research for this thesis, while ShapeSpace went about assembling one cohesive master list from those that had been provided, network theory was used to build an accurate model of project A from the same lists. This network formed the basis of the new master list and was used to underpin a graph database. This graph database was then employed to query and test the data, allowing the collection provided by company A to be assessed and consolidated.

Interestingly company A estimated that 117 new BOMs had been created in the last 3 years, but initial investigations showed that over 850 had been created in that time period. This was due to company practice, where engineers would create a new BOM for every customer who requested a new product. This is a notable, current example of a common issue within industry, where companies cannot accurately understand their data through the PLM systems they employ, as suggested by Myer [138] and others discussed in subsection 2.5.3. Within company A, instead of modifying a modular component to build a new specialised product, a designer would take previously created models that were similar to what the customer had requested and modify them, then save them as a new product. While this in part complies with desired goals for design reuse, it meant there was a proliferation of new BOMs and the database had become very large and complicated.

At the same time as ShapeSpace were working on the creation of the master BOM list, removing duplicates from the multiple lists and providing clear material information for each CAD file, this research investigated the underlying network. It was thought likely that network theory could provide methods to assist the work ShapeSpace were doing. Moreover, accomplishing this would allow company A to assess their current components and consolidate some assemblies to make them modular and customisable without having to manufacture totally original products for each customer. Using network theory it was possible to build a multimodal network of the data and investigate it, potentially allowing new insights into company A's data.

6.3.1 Methodology

The BOMs from company A were stored in spreadsheets, where each BOM was a tree list; assemblies containing sub-assemblies, sub-assemblies containing components. Alongside the BOM data, order data was provided. The several lists provided were transformed into a single, complete network. The network included links between items and their original source spreadsheet, this information was preserved for ease of processing. This network was the basis of a graph database, created using TinkerPop, to allow access and traversal of the data. As an initial step in investigating the data structure of company A, a simple multimodal graph of the data was created and is shown in figure 6.3.

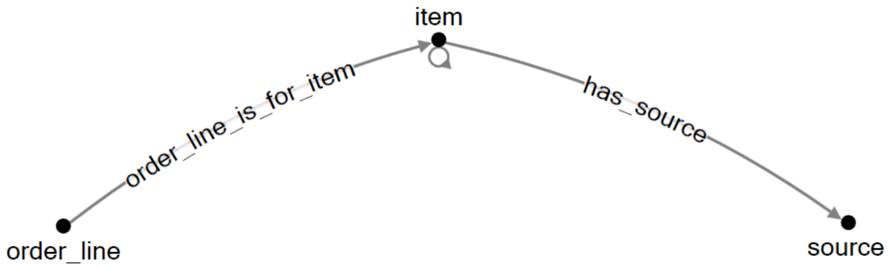


Figure 6.3: Network 9.1: Initial network for company A

Figure 6.3 is a small directed multimodal network, showing the types of data from company A on the nodes and the relationships on the edges. In network 9.1 'source' refers to the spreadsheets where the data originated and the self-loop edge to the 'item' node is present because of the relationship 'item contains item' between these types of node, called 'contains' elsewhere in this work. This simple network allows a clear overview of the data provided by company A and corresponds to the expected structure. 'Item' information was stored within several source spreadsheets, while the 'order_line' information was provided separately. Building a simple network, such as network 9.1, allows basic checks to be performed on the data and ensures that it is correctly interpreted. Using network visualisation to map a company's CAD data structure, such as figure 6.3, allows a clear interpretation and overview of a collection of CAD and the other data to which it is linked.

It is interesting to note that network 5 (figure 4.7), a directed bimodal network presented in section 4.2, is the data represented by the middle top node labelled 'item' in figure 6.3. Considering the complexity of network 5 and how it relates to network 9.1 highlights the importance of choosing which elements from a company's CAD collection to model. It would be possible to construct a new network, expanding network 9.1 to show the level of detail in network 5, resulting in a very large, complex network of all the data provided by company A. In that network 'item' nodes would be linked individually to the order lines travelling to them and to the sources they come from, instead of by trunk edges as shown in figure 6.3. This complex network of project A from company A would allow a measure of diameter and density to be calculated.

Utilising the TinkerPop graph database capabilities it was possible to assess the network built from company A's CAD collection, based upon network theory. It was decided to explore the BOMs and components, and an attempt was made to count how many items each BOM contained and how many BOMs each item was contained in. Gremlin was used to return queries from the database and these were used to plot the results. Exploration of BOMs and items also compared statistics for the total number of items in the system and for the leaves of the assembly (the individual components at the bottom of the assemblies that are contained within several assemblies, but which are single components). Simple line plots were built from the results of these queries but first the predicted results are shown in figure 6.4.

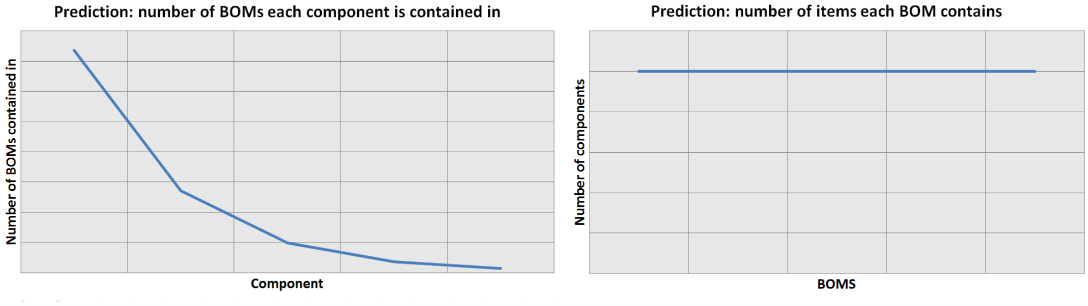
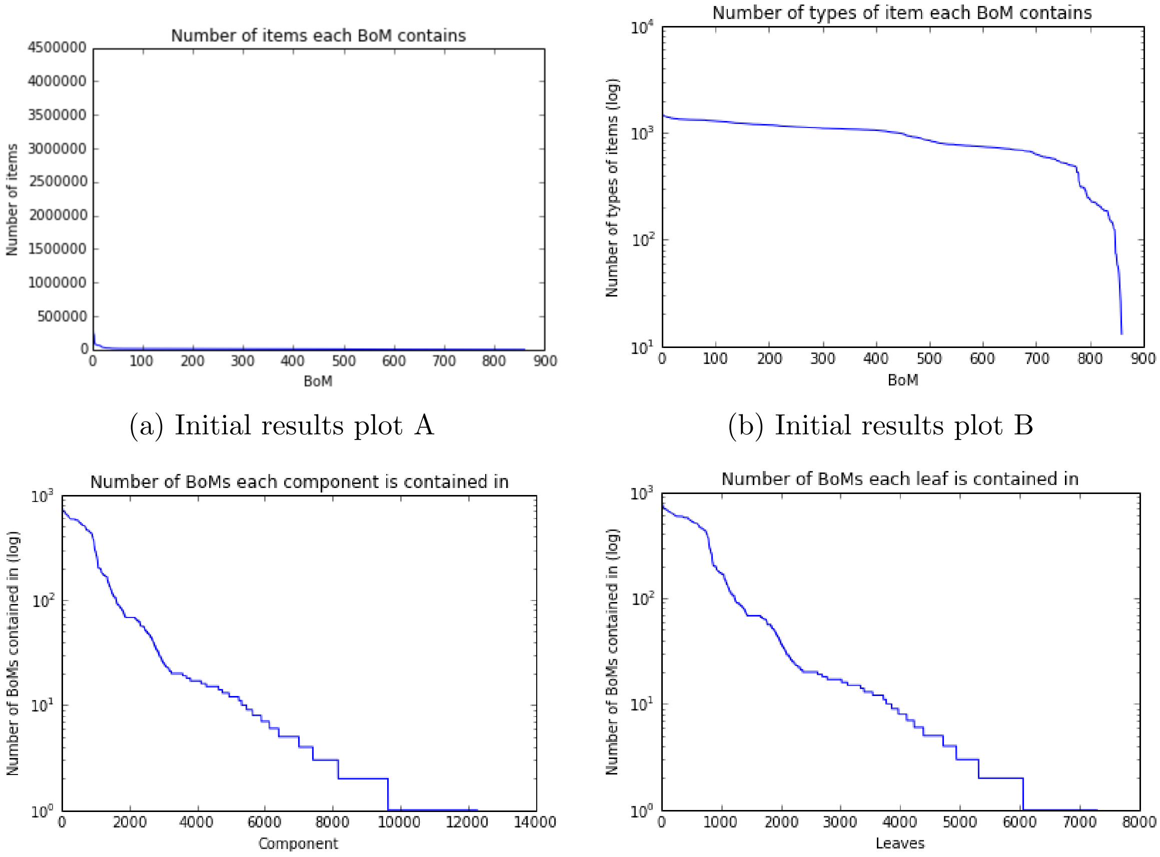


Figure 6.4: Expected plot shapes of company A's data

It was expected that the results of how many BOMs each item was contained in would be represented by an exponential decay, depicted by the left-hand plot in figure 6.4, as components such as screws, nuts and bolts would be commonly used while other leaf components would be more specialised and used less. It was supposed that the results of how many items each BOM contained would be relatively flat, illustrated by the right-hand graph in figure 6.4, as each final product, being of the same series in project A, would contain roughly the same number of parts.

In order to generate these plots from the graph database queries were written, focussing on the nodes of type ‘item’ only, to find the leaves and roots of the BOMs. Initially the approach was to loop through the paths of the graph, identifying each leaf and working backwards, up from there to the root (top-assembly) of the BOM. This was deemed too computationally expensive, due to the high number of CAD files. Instead a query was written that would loop from root to leaf necessitating the identification and storage of all the roots. This stored list of top-assemblies was then used to begin the query loop, which was programmed to stop after 20 iterations and store all the leaf nodes found. The results of this query were ordered to find a count of how many items were in each BOM and how many BOMs each component was contained in, and the results are presented in figure 6.5.

6.3.2 Initial results



(a) Initial results plot A

(b) Initial results plot B

(c) Initial results plot C

(d) Initial results plot D

Figure 6.5: Plots of initial results from company A’s data

The plots in figure 6.5 show that the results were as predicted; the shapes match those in figure 6.4, with most BOMs containing a similar number of items (figure 6.5(a) and (b)) and the items themselves being contained in a range of BOMs, following an exponential decay curve (figure 6.5(c) and (d)). There was one outlying result in figure 6.5(a), which appeared to be throwing the rest of the graph off. In addition the axes on the plots appeared incorrect, as they were displaying much larger scales than expected.

Assuming there were errors in the original queries, these were rewritten and run several times more. However, these further iterations made it clear there were no errors with the original queries and, consequently, the result plots in figure 6.5 were shown the be correct, and there was no error with the axes. Further assessment showed that the results of the queries, prior to the plots being created, did not seem logical; One BOM was registering as having over four million sub-assemblies and parts (the outlier in figure 6.5(a), forcing the scale to be so large), which was known to be false and failed simple 'common sense' tests about the data supplied by company A. Investigation continued, and the underlying network for the graph database was rebuilt, however the same results as shown in figure 6.5 were returned. The persistent error highlighted that the queries run on the graph database were correct, but there was a flaw in the underlying network. It emerged that the network had been constructed incorrectly from the spreadsheets provided by company A, meaning that all queries and plots were giving false results.

6.3.3 Error in network creation

The network that was being used, underlying the graph database that had been queried, was incorrect due to a bug in the original code used to create it. It is unlikely this error would have been exposed without the network analysis that was performed on the data. As all the data had been provided in list form, python code was written to read this and produce a network. These lists, in spreadsheets were in a format similar to the example in table 6.1

The network that should result from this particular list is shown in figure 6.6, but the network produced was like that shown in figure 6.7. Both network diagrams were arranged by hand to clearly illustrate the levels involved in the BOM example from table 6.1. The key difference is the way the relationships, appearing in list for in table 6.1, are recorded. In figure 6.6 if a relationship has been counted more than once from table 6.1, the value is given to the link between two nodes, most commonly as a weight given to the edge. This weight, displayed in figure 6.6, is shown by a heavier edge connecting two nodes, such as sub-assembly b and sub-assembly c on the left-hand side of the diagram, showing clearly that sub-assembly b contains more than one sub-assembly c, in this context.

Level	Item
0	A
1	B
2	C
3	D
3	D
3	E
2	C
3	D
3	D
3	E
1	F
2	G
3	H
3	H
1	I
2	J

Table 6.1: Example of BOM data list format

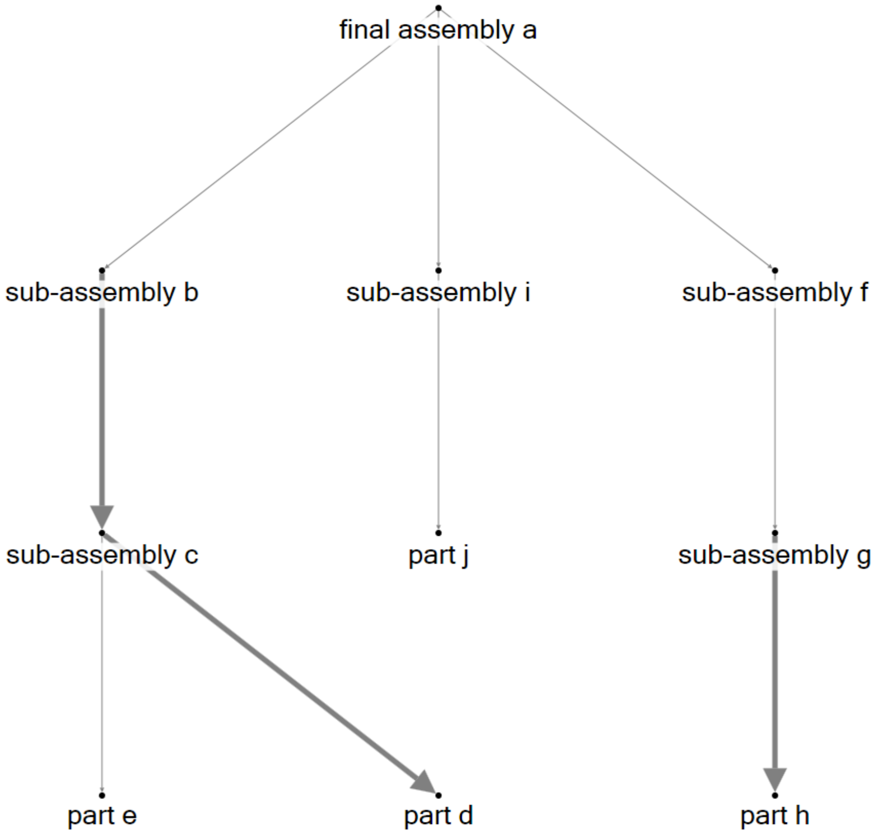


Figure 6.6: Example of correctly built network graph from table 6.1

While it can be argued that the network diagram in figure 6.7 also shows that sub-assembly b contains two of sub-assembly c, shown on the left-hand side of the diagram, these 'extra' edges cause problems when looking at the network as a whole. The incorrectly built network in figure 6.7 suggests there are two entries for sub-assembly c, four occurrences of part d and two occurrences of part e in the network, which is false. While the final assembly 'A' contains than number of those items, the data the BOM is built from does not. If a query were to be run comparing the number of vertices contains within the networks shown in figures 6.6 and 6.7, figure 6.6 would return a value of 4, while figure 6.7 would return a value of 9. This simple example illustrates how the global metrics, and node level metrics, of the resultant network are affected by the incorrect build.

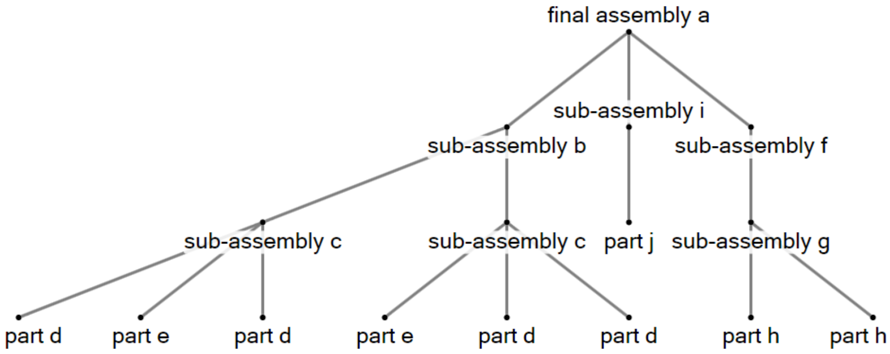


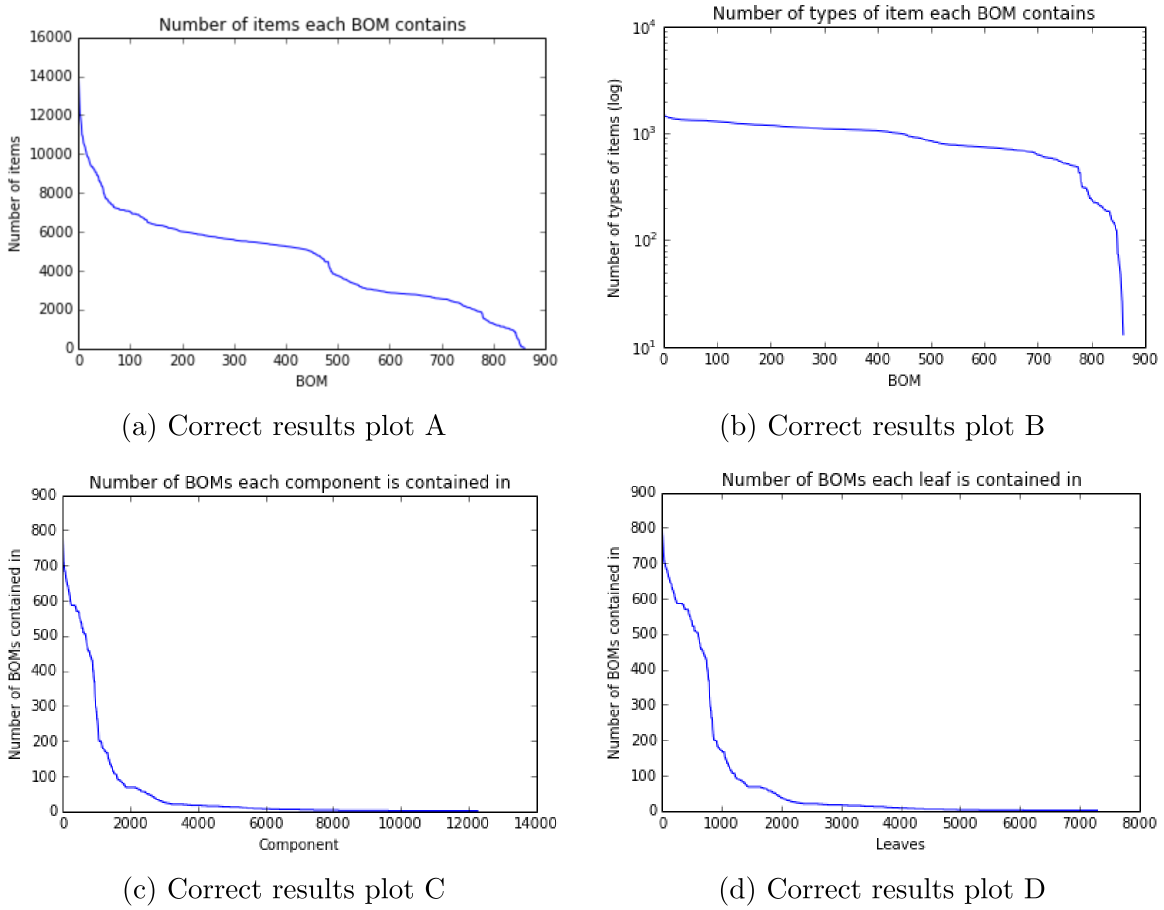
Figure 6.7: Example of incorrectly built network graph from table 6.1

This example illustrated the type of problem encountered when building the initial network from company A's data, where the network constructed by the code was found to be incorrect. As this was the source of the error, the network underlying the graph database had to be rebuilt before research could continue.The bug in the code was corrected, so duplication of relationships and multiple occurances of the same parts were avoided.

6.3.4 Correct network results

With the base network remade and examined for accuracy, investigation into the data was continued. The queries were re-run and the results plots remade, shown in figure 6.8.

The plot of number of items each BOM contains, shown in figure 6.8(a), follow a gradual decline rather than the predicted straight line, while the number of types of item each BOM contains, shown in figure 6.8(b) is a relatively straight line, similar to



(a) Correct results plot A

(b) Correct results plot B

(c) Correct results plot C

(d) Correct results plot D

Figure 6.8: Correct result plots from company A's data

the predicted results in figure 6.4 illustrates. Figure 6.8(c) and (d) show the plots for number of BOMs each component is contained in and the number of BOMs each leaf is contained in respectively; both are exponential decay curves, matching the expected results from figure 6.4.

6.3.5 Further mapping: Network 9.2

Network 5 is visualised in figure 4.7 and was created by capturing all the 'item contains item' relationships and using them to create a network separate from the rest of company A's database. The nodes represented various CAD files and the links between them were 'contains' relationships. This was a very large CAD collection and the impressive visualisation illustrated its complexity, especially in comparison to smaller collections such as that modelled in network 3.1. While the metrics of this network were discussed in chapter 4, it is notable that these are calculated using the CAD collection from the whole data structure represented in network 9.1 (see figure 6.3). It is expected that these metrics would change significantly if further data was added to the network, expanding it from a bimodal model of company A's CAD collection to a multimodal modal of the whole data structure.

Expanding network 9.1 (figure 6.3) by including more data types from the whole of company A could create an accurate whole network model of company A. This would be possible using the network that was built to underpin the graph database used in section 6.3. It would be expected that, as the number of nodes would be high, the number of edges would also be high and the diameter of the network would be relatively small, though larger than the reported diameter of network 5. In network 9.1 (figure 6.3) the types of data are shown by summarised nodes and the diameter of the network is reported to be 4. It is expected that expanding network 9.1 would also increase diameter, especially if summary nodes were expanded from types down to individual 'items'. The graph density would be expected to decrease, as the data being modelled by this expanded network is not randomly assigned, unlike networks 3.2 and 3.3 presented in section 4.4. The relationships being modelled are already assigned and are not likely to have many links to many adjacent nodes, as would be found in a dense network. It is not simple to predict the changes in node specific metrics, however it would be expected that any CAD data modelled would be most central in a network modelling the whole company's data structure with large in-degree metrics, while orders would have large out-degree measures highlighting their influence.

In order to create a network diagram of the expanded data structure, a large number of relationships were taken at random from the graph database used in section 6.3. Due to the very large size of the whole database it was considered reasonable to use a portion of it as a representative model, allowing more straightforward and rapid computations as well as giving a partial model to test the expected results. As such 15000 entries were randomly taken from the original database and NodeXL was used to create figure 6.9.

Network 9.2 is a multimodal directed network presented in a Fruchterman-Reingold layout. It is notable that despite the vast number of relationships taken from the network used in subsection 6.3.1 this mapping actually resulted in network 9.2 being markedly smaller than network 5. The metric for edge weight was mapped to edge width and directed edges are indicated by arrows, where heavy edges indicate a strong link or a large number of 'contains' or 'order is for' links. It is unlikely these heavy edges

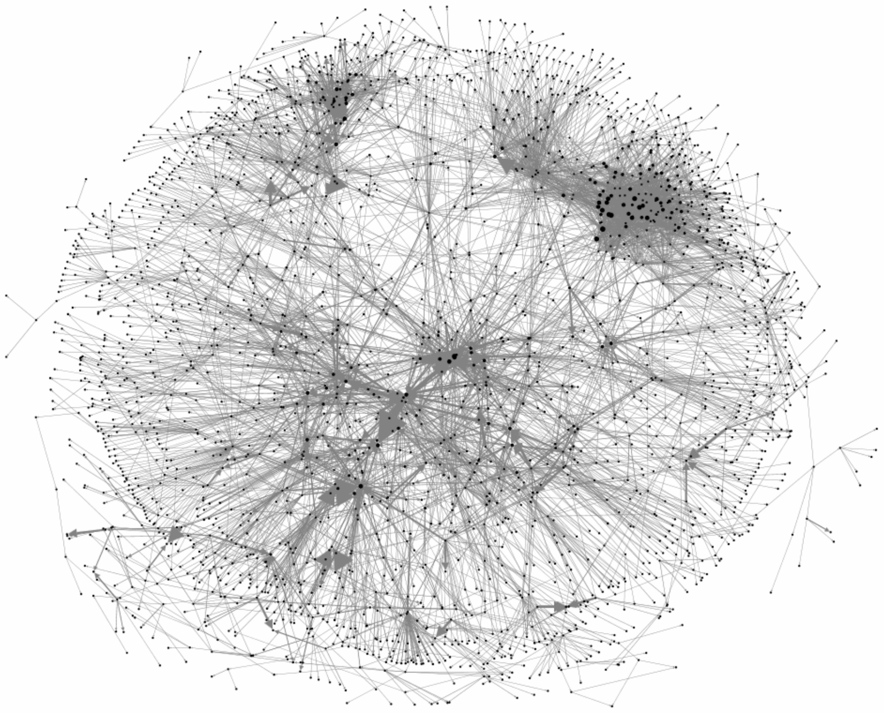


Figure 6.9: Network 9.2: Partial model of the data structure of company A

would show a relationship between an item and its source, as each item is presumed to have a one-to-one relationship with its source. The other metric that is applied to a visual property is in-degree, which is mapped to node size. Some key metrics to allow comparison are presented in table 6.2.

Metric	**Network 9.2**
Type of graph	Directed
Number of nodes	2419
Number of edges	5623
Number of self-loops	0
Maximum geodesic distance (Diameter)	14
Average geodesic distance	5.6
Graph density	0.0019
Minimum clustering coefficient	0
Maximum clustering coefficient	1.0
Average clustering coefficient	0.002

Table 6.2: Metrics of network 9.2

Network 9.2 is reported to contain 2419 nodes and 5623 edges. This is smaller than was expected from the data collected; it was expected that the 15000 lines read out of the database would result in a network with 15000 edges, however it appears that many of these data lines were duplicates, and therefore the edge numbers were counted and consolidated into edge weights. This indicates that the network underlying the graph database has a considerably different structure to the resulting network 9.2 in figure 6.9. There are no self-loops within the data sample, which is as would be expected as no individual item should be linked to itself, no order should contain itself and no source would be linked back to itself. The diameter of the network is reported to be 14, which is comparable to that of network 5, despite the difference in the network sizes. This could be due to the different types of data that multimodal network 9.2 models, unlike the bimodal data that network 5 models as well as the partial data mapped. If the network was complete the diameter might be different. The graph density is also low, though higher than that reported for network 5. This may indicate that a network comprising a company's whole data is more connected than one showing only CAD data.

The 15000 entries from the database that were used to build network 9.2 were taken randomly from over 200000 original data-lines, using a Gremlin query of the graph database. The random selection of these nodes is not directly comparable to the random networks (networks 3.2 and 3.3, figures 4.8 and 4.9) in section 4.4. While the edges themselves were randomly selected from the whole database, the relationships they model are not random, but are meaningful data representing the structure of project A. As much of the data explored prior to this did not exhibit characteristics commonly described as 'small world' features (measured by the clustering coefficient metric), it was thought this large company data structure would also not exhibit small world structure, despite having a larger number of connected data types. The clustering coefficient values appeared to confirm this, as shown in table 6.2. The results of this further mapping indicate that continuing this investigation and expanding the network to fully model the database of company A could return more interesting results, especially when considering a company's whole data structure.

6.3.6 Complete mapping: Network 9.3

Continuing this investigation, the whole of the graph database was transformed into a network, presented in figure 6.10, network 9.3.

Figure 6.10 shows network 9.3, a directed multimodal network which was built from the whole of company A's data. Each relationship line was read out of the graph database and transformed into a network using NodeXL. The network was laid out using the Fruchterman-Reingold layout with the arrows removed from the directed edges for a simplified view. The network is exceptionally large, containing 17306 nodes and 201840

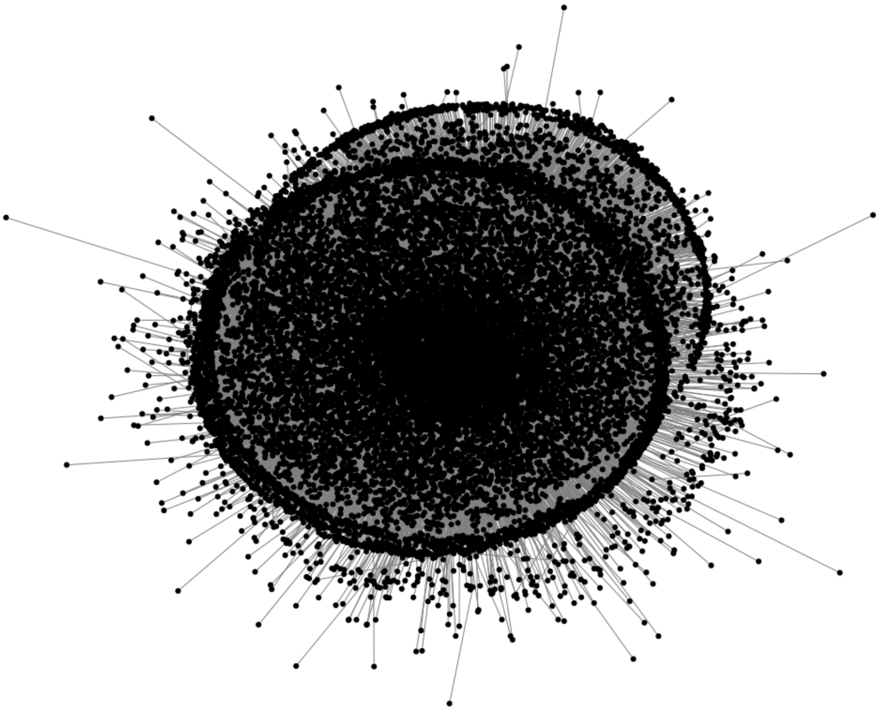


Figure 6.10: Network 9.3: Complete model of the data structure of company A

edges and was computationally expensive, taking over 12 hours to build. The many nodes are clustered together, often overlapping so that they are difficult to distinguish individually. Some edges can be clearly seen, though the majority overlap and are also difficult to differentiate. It may be possible to improve this using metrics mapped to visual properties. Table 6.3 presents the metrics for network 9.3

Metric	**Network 9.3**
Type of graph	Directed
Number of nodes	17306
Number of edges	201840
Number of self-loops	0
Maximum geodesic distance (Diameter)	8
Average geodesic distance	3.03
Graph density	0.00055

Table 6.3: Metrics of network 9.3

The metrics calculated and shown in table 6.3 present the huge size of network 9.3, with the nodes representing three data types and the edges showing three relationships. As expected there are no self-loops within the network, indicating that it is completely connected. It was supposed that adding all the nodes to network 5 or 9.2 would increase the diameter, but the results here show the diameter of this network to be 8. While this is larger than Milgram's supposed measure of 6 [130] it is relatively low when considering the size of this network, suggesting that adding nodes to create a multimodal network of whole company data may create a more useful model of CAD collections. These additional nodes are not entered into the network randomly but with meaning and may provide a more searchable network, where CAD models are more closely linked. This measure is much reduced from that of network 5, which modelled only company A's CAD data, shown in figure 4.7. Also of interest is the graph density, which is an exceptionally low value, suggesting real world CAD and associated data collections may be sparse data structures. These features suggest that a network model of an industrial CAD collection and associated data are small world networks, introduced by Watts and Strogatz [197]. It might also be considered that this diagram and associated metrics illustrates CAD file structure is a scale free network, where hubs can be identified linking many other nodes, however this requires further investigation.

These measures suggest that a complete network model of a CAD collection and associated data provides a network that is quite easy to traverse, indicated by the relatively small diameter of the network. Also the density indicates that the network is sparsely connected, suggesting that CAD collections are denser when considered separately from associated industrial data. This is a significant model of a real world data collection from a mechanical engineering design and manufacturer. The network, visualisation and calculated metrics are novel and suggest network methods could be effective in categorising company design data, as well as assisting in search methods.

6.3.7 Company A data structure discussion

The correct result plots from the analysis of company A's data, shown in figure 6.8, agreed with the predicted results, in figure 6.4. It was expected that the number of items each BOM contained would be similar, however there is a variation in the number of items each BOM contained, signifying that there are some BOMs that are more complex than others. Despite this, the majority of BOMs contained a similar number of types of item, revealing that few final products are complex in type. The number of BOMs each component is contained in followed the predicted exponential decay curve, as did the number of BOMs each leaf is contained in. This shows that some parts are more commonly used than others.

Further analysis of these results uncovered pertinent details. In the whole project, there were 80 duplicate top level assemblies (products), with several being identical in every respect yet having their own BOMs. This is an example of unnecessary proliferation. There were also over 250 duplicate assemblies. Other interesting statistics include:

- 16% of components were the most commonly used and appeared in over 100 final products
- 47% of components appeared in fewer than ten final products
- Over 1200 components (16%) appeared in only one top level assembly (product)
- Five products contain over 130 component types while the majority (>90%) contain less than 40 types
- Five products contain over 1700 components while 97% contain less than 100 components

These statistics demonstrate the complex the CAD collection and also highlight how improvements or optimisations of some products and components could greatly affect the collection. Grouping top level assemblies (products) by similar components would create groups, showing company A where modularisation would be most effective.

When these results were presented to company A, interest was shown. It was seen as valuable that engineers within the company could access information on product volumes. However the proliferation of parts within the project and the vast quantities of inactive data did not concern the company, as they claimed many of the duplicated BOMs and assemblies would be considered 'not live' within the company's design. While this analysis was well received by company A, these results were not taken further or acted upon due to a number of factors, including economic issues.

Using network theory allowed errors in the initial analysis to be found and statistics about the projects BOMs to be easily identified, due to the structure provided. It would be possible for a company such as ShapeSpace to use this methodology to perform similar analysis on other company data. It is interesting however that the real world data here gave rise to unexpected problems, indicating that there is no one way to do things in a manufacturing context. As with the initial incorrect results returned, real world data may be incomplete or contain mistakes such as incomplete metadata, which would affect results.

While it may have been possible to assess the data in other ways to discover these statistics, modelling the data using a network database that could be queried has been shown to be a robust and reliable method for this analysis. While the initial results shown by plots mirrored the expected results, it became clear there was an error, which was discovered to be with the initial network build. Using another technique for this analysis, such as a relational database, may not not have shown that the initial build contained an error, and any further results would have contained errors. Network theory provided a novel way to model company A's data and produce plots of statistics that interested them.

When comparing the data structure of company A (figure 6.3) to the predicted structure (figure 6.1), it is clear that network 9.1 is only part of the expected structure. Network 9.1 clearly exhibits the item and order (order line) nodes that were predicted but does not contain any other predicted nodes. This is due to company A only providing CAD and order data for analysis. Also the node labelled 'source' was not expected, however it can be seen that this node is due to the data analysis performed to form the network and as such, is not a feature of the data structure of company A. This simple data structure partially agrees with the predicted data structure of industrial CAD collections.

The further mapping of 15000 lines from the graph database, expanded this simple data structure visualisation, but only partially. The 15000 relationships, chosen at random, were turned into edges in network 9.2, shown in figure 6.9. Comparing the metrics of network 9.2 to those reported for network 5, an entire mapping of company A's CAD collection indicated that the multimodal network, though smaller in node and edge count, has a larger diameter. This suggested that, while network methods offer an interesting and novel way to explore a company's CAD collection, multimodal networks of the whole data structure would result in significantly different metrics. The model of the whole of company A's data structure, network 9.3 (figure 6.10) shows how network metrics and visualisations are significant when considering design data. These metrics could be used to indicated characteristics of a company's whole data structure and it is suggested this work be continued.

6.4 Company B Data Structure Investigation: Networks 10.1 and 10.2

Company B is a large engineering manufacturing company, specialising in products and services for oil, gas and power industries. Company B had engaged ShapeSpace in a consolidation project similar to that performed for company A, on one of the company's large manufacturing areas, which for the purposes of this work shall be referred to as project B. It was decided that the investigation of the data provided by company B should explore different aspects of network theory, not mirroring the analysis performed on company A's data. It was thought that further aspects of network theory could be explored by using different methodologies on another large real world CAD collection.

To begin the investigation the data from this industrial CAD collection was used to build a network, providing an overview of the data structure, which could be compared to the predicted results presented in section 6.2 (figure 6.4). From there it would be possible to further explore the CAD collection and associated data.

6.4.1 Initial mapping: Network 10.1

Instead of creating a network to underpin a graph database as had been done for company A, NodeXL and network methods were used to construct a simple network of company B's data structure. Network 10.1 is a multimodal directed network where the nodes represent different types of data, not only CAD data or individual CAD files, and the edges show the relationships between the data types. This is in contrast to the networks previously built (networks 1-8), but similar to network 9.1 (figure 6.3). Network 10.1 is shown in figure 6.11 and was arranged by hand to allow a clear layout, while figure 6.12 shows the same layout with quantities displayed.

Network 10.1 was built by identifying the different types of data contained with company B's database and assigning each to a node. These types were then counted and the values assigned to the corresponding node. The edges are labelled to allow clear identification of relationships between nodes, shown in figure 6.11, while the number of these links were counted and assigned to the corresponding edge. Quantitative data is displayed in figure 6.12 for clarity. Building this network was straightforward from company B's data and enabled quantities to be checked before further analysis while providing an image of the structure of the data. The metrics of network 10.1 are shown in table 6.4.

The directed network clearly illustrates the relationships between the different categories of data from company B. The node labelled 'order' at the top centre of the diagram has two edges directed away from it, towards 'customer' and 'order_line'. These relationships are labelled on the edges and reveal this to be an influential node with the largest out-degree in the network. The node in the bottom right of the diagram, labelled

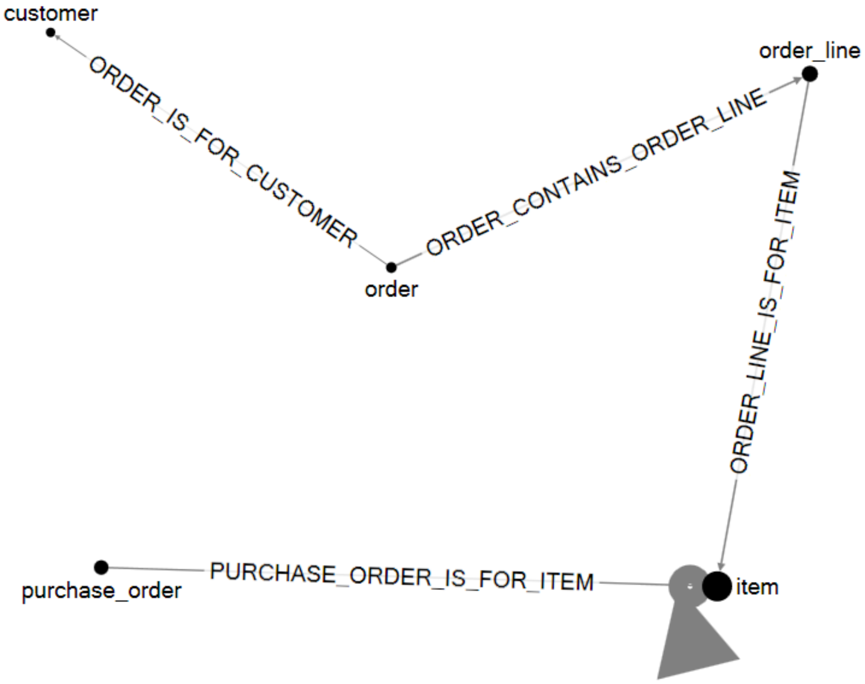


Figure 6.11: Network 10.1: Initial network for company B

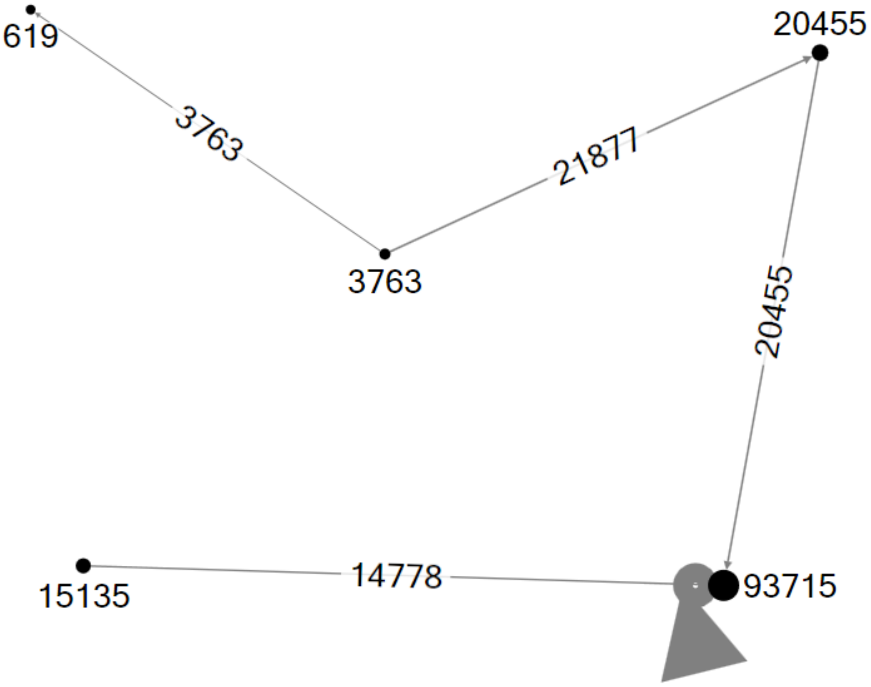


Figure 6.12: Network 10.1: Initial network for company B with quantities

‘item’ is the node representing the CAD data, and it has two edges travelling to it from the nodes labelled ‘purchase order’ and ‘order_line’, so it has the largest in-degree in the network. It also has a very large self-loop, shown by the large grey circle and triangle, representing the ‘item contains item’ relationship. While in this network diagram that relationship is a self-loop, it is equivalent to the ‘contains’ relationships modelled in networks 1, 2, 3.1, and 5. As such, it indicates how many times individual CAD files are used within company B’s collection and, when compared to the number of nodes, may give an indication of how well used the files are within the CAD collection.

Figure 6.12 presents the quantities for each node type and relationship. The 3763 orders influence the 93715 items, via 20455 order lines. Accordingly the 3763 orders are linked 21877 times to the order lines as some will be duplicated. In a similar way it can be seen that the order amount is not equal to the number of customers (619) as each customer will place multiple orders. The value at the purchase order node shows how many purchase orders were placed for items that needed to be bought in by company B. The difference in the value on edge ‘purchase order is for item’ at the bottom of 6.12 and the purchase order node could be due to multiple purchase orders being made for the same items.

Metric	**Network 10.1**
Type of graph	Directed
Number of nodes	5
Number of edges	5
Number of self-loops	1
Maximum geodesic distance (Diameter)	4
Average geodesic distance	1.6
Graph density	0.2
Separate components	1
Highest In-degree	item node (3)
Highest Out-degree	order node (2)
Heaviest Edge	item contains item (456438)

Table 6.4: Metrics of network 10.1

The diameter of the network is just 4, which may be gauged from the diagrams in figures 6.11 and 6.12, while the average geodesic distance is found to be 1.6 and the graph density is 0.2. If the network was to be expanded, rather than summary nodes representing types of data from company B, nodes would represent individual CAD files, orders, customers and purchases, and it would be expected that these global metrics would change. If all CAD files were linked to a purchase order, and to a customer through an order and order line, the diameter of the network could remain the same.

The metrics of in-degree and out-degree show the most influenced and most influential data in the network respectively. In network 10.1 these are, inevitably in this context, the item and order nodes. Within industry it would be assumed that orders are the data that most affect a company's efforts, while their designs are the most influenced data, driven by these orders. It is evident that the CAD data is the largest portion of the network, shown by the reported metrics as well as by the large node and large self-loop in the bottom right of both figures 6.11 and 6.12.

6.4.2 Further mapping: Network 10.2

To further explore the data, network 10.1 was expanded to include more detail about the 'item' type data. The types of item were counted and included as individual nodes, in place of the one 'item' node. Figure 6.13 shows the resulting network diagram of this advanced mapping.

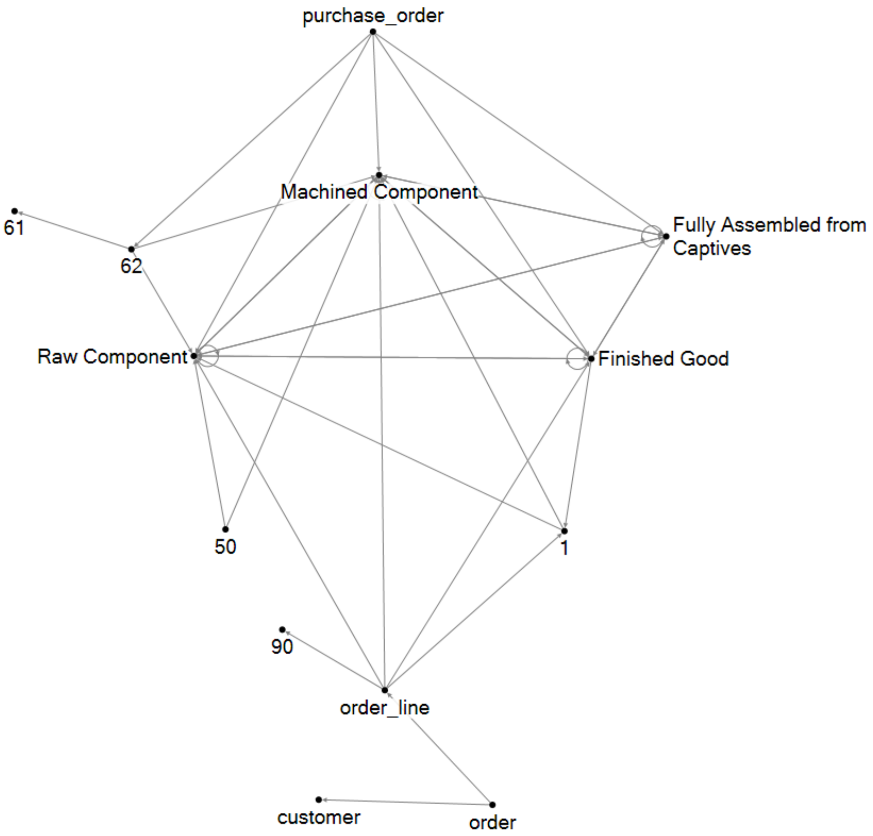


Figure 6.13: Network 10.2: Initial mapping of CAD item data

Network 10.2 is a multimodal directed network, where the bottom right node 'item' and self-loop edge of network 10.1 have been expanded to show the relationships between the types of item. In network 10.2 there are several nodes which represent item types and the edges between them show 'item contains item' relationships. All items have numerical identifiers, shown in figures 6.14 and 6.15, but in figure 6.13 the top 4 item types are named to allow clear discussion. These key nodes are detailed in table 6.5 for clarity.

Numerical label	Item Type	Category size	In-degree	Eigenvector Centrality
20	Machined Component	20644	9	0.153
30	Raw Component	29173	9	0.153
2	Finished Good	26464	6	0.134
31	Fully Assembled from Captives	8635	5	0.105

Table 6.5: Key nodes in network 10.2

These four nodes, displayed along with some key metrics in table 6.5, were identified as important by their metrics and the number of individual files they represented. Of the nine item categories these were the largest four, containing over 90% of all the items in company B's CAD collection. Notably they also exhibited the largest values for in-degree and eigenvector centrality. The high value for eigenvector centrality shows that in the network these nodes are the most influential, while the high in-degree points to many other items relying on them, or orders or purchases being linked to them.

Figure 6.13 shows the relationships between the different types of CAD data, clearly mapping which items rely on others. There are some self-loops, not because any individual items within an item type category contain themselves, but because individual items contain other singular items within the same type category. This data structure is not similar to an assembly structure, like that shown in networks 1, 2 and 3.1; instead it is much more like a BOM structure, where CAD files are linked to those they rely on. It is clear that there are purchase orders for certain types of items only, but there are purchase orders for 'finished goods' meaning company B may buy in some items to resell. Also notable are the order lines, which do not only travel to 'finished goods'. This may mean company B are selling incomplete products or sub-assembly items as spares, or that the items are being sent to another part of the company outside of project B.

To further investigate this data figures 6.14 and 6.15 were produced. These images show network 10.2 with different metrics mapped to edge weights.

In figure 6.14 network 10.2 has the purchase order values mapped to edge width. This diagram clearly shows how the purchase orders are divided between item types. It is apparent that the most commonly purchased items are machined components, then raw components and fully assembled from captives items. It would be expected

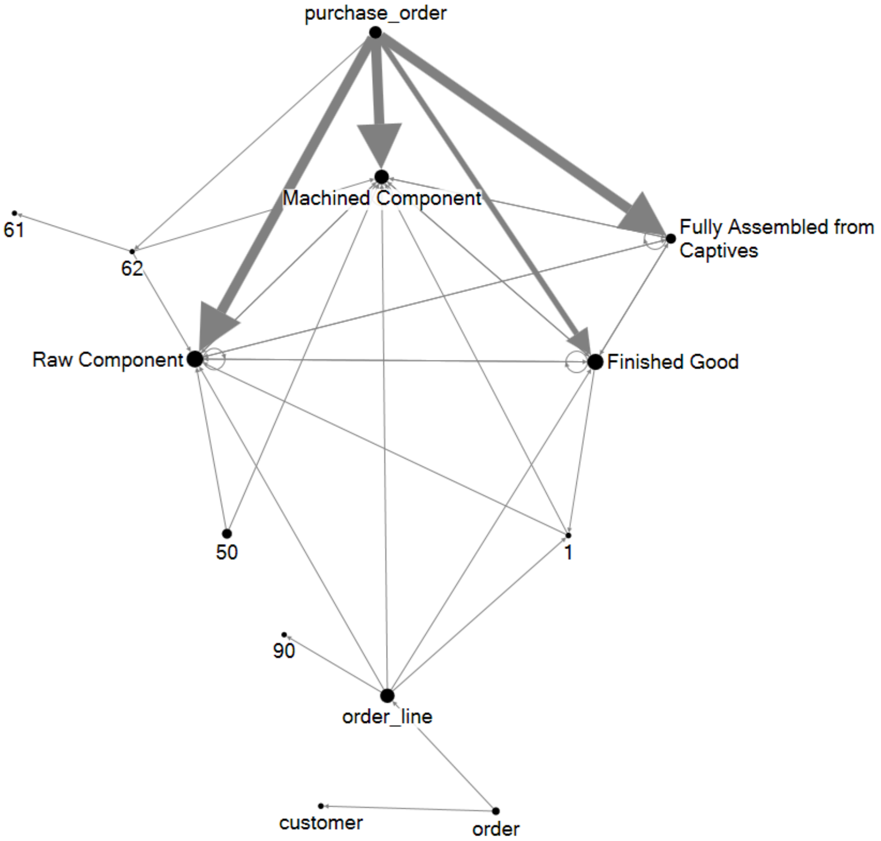


Figure 6.14: Network 10.2: Purchase order values mapped to edge widths

that an industrial manufacturer might purchase some simple raw components and then transform or develop them, especially when manufacturing such items from scratch on site may be more costly. The network diagram in figure 6.14 could be instrumental in allowing company B to assess the purchase orders they were placing and assess where to focus efforts in order to reduce costs.

In company B, it appeared that many of the raw components they purchased underwent further processes to create a finished item ready for sale and this was clearly illustrated by another metric mapping, shown in figure 6.15. This figure is another representation of network 10.2 but with different edge weights mapped to edge widths. Edge weights have been calculated and applied only to ‘item contains item’ edges for clarity, based on the number of edges between the item types. It can be seen that most item types have similar sized edges travelling between them, however there are two particularly large edges. These heavy edges show finished good items rely mainly on raw components and machined components, while there appears to be an equal distribution of use among the other item types. These strong links could indicate where

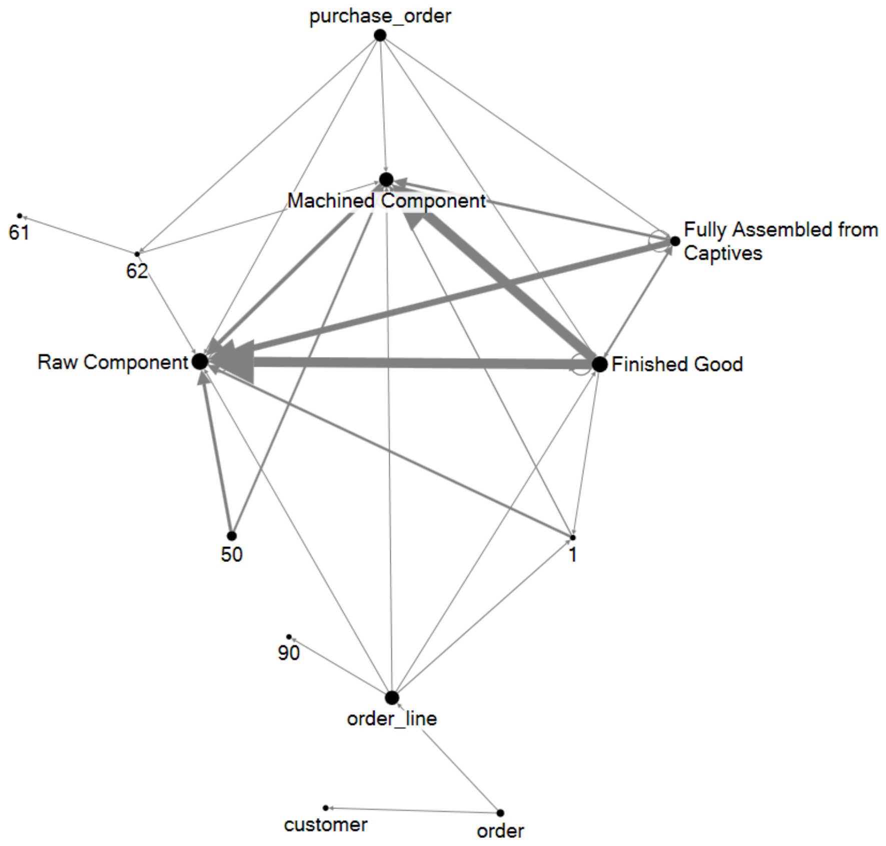


Figure 6.15: Network 10.2: Edge width representing edge weight within CAD data

the company could make improvement to designs. Also the node types that are less connected, labelled '50', '61' and '90', could indicate specialist parts or parts that have become unused.

Figures 6.14 and 6.15 show interesting diagrams of the same data, with different metrics emphasised, illustrating the value in mapping data values to visual metrics in a network. Both diagrams provide clear visual information on the strength of link between item types or purchase orders for item. While the data modelled appears visually different in the presented diagrams, the global metrics of network 10.2 remain the same and are presented in table 6.6.

In table 6.6 some of the global metrics for network 10.2 are recorded. There are 14 nodes and 37 edges with 4 self-loops. These self-loops are all attached to item type nodes, suggesting they are there as different individual items may contain another item from within the same category. Comparing these metrics to those presented in table 6.4 shows the difference between the two representations of company B's data. The number of nodes and edges has increased as would be expected, and the number of

Metric	**Network 10.2**
Type of graph	Directed
Number of nodes	14
Number of edges	37
Number of self-loops	4
Maximum geodesic distance (Diameter)	5
Average geodesic distance	1.9
Graph density	0.18
Separate components	1
Highest in-degree	Machined/ Raw Component (9)
Highest out-degree	order_line (6)
Heaviest edge	Finished Good contains Raw Component (269043)

Table 6.6: Metrics of network 10.2

self-loops has also increased. This is understandable, as some of the items will contain the same type as themselves. The diameter of the network has increased from 4 to 5 but the graph densities are alike (0.2 for network 10.1 and 0.18 for network 10.2). Despite increasing the number of nodes and edges, the similarities of the diameter and density measures mean the networks are comparable sizes. This could indicate that CAD data is scale free, but further work would be required to conclude this. The highest in-degree measures now correspond to the machined and raw component nodes, suggesting they are the most influenced data in the network, but also indicating they are the most relied upon. The highest out-degree measure is for order lines rather than orders now, mainly due to the expansion of the item node into many type nodes, but still indicates that orders are the most influential data in this collection. The heaviest edge is now located between the finished good node and the raw component node, as seen in figure 6.15, demonstrating the key nature of this relationship. This investigation was continued by different visualisations of network 10.2.

6.4.3 Further visualisation of network 10.2

The visualisations presented in subsection 6.4.2 indicated that it may be possible to represent the structure of the company data clearly using network diagrams, however only basic visualisations were presented. This work continued to explore the visualisation possibilities and the resultant mappings are presented. The network diagram from figure 6.13 was modified and different metrics represented as visual properties to assess their use. Figure 6.16 shows the first visualisation produced in this next phase.

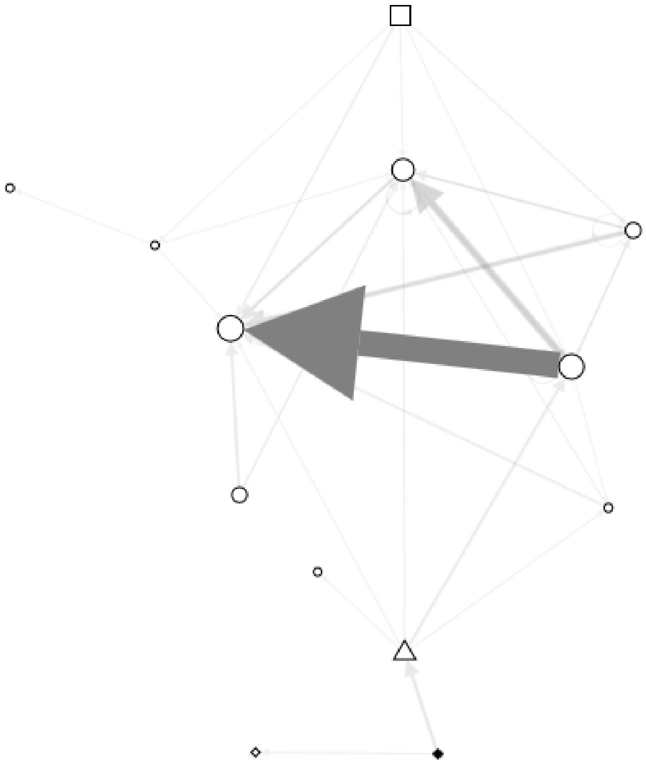


Figure 6.16: Network 10.2 with mapped edge weights

In figure 6.16 all edge weights have been used to apply a width to each edge. This has resulted in one very large edge travelling between two nodes in the centre of the diagram, obviously the largest link in the network. From earlier diagrams it is know that this edge travels from the 'finished good' node to the 'raw component' node but, if this prior knowledge was unknown, this heavy edge would indicate the key nature of this link in the network. Differing shapes have replaced the circular nodes to allow node type to be distinguished. While these shapes do not divulge the data they represent, they do illustrate the different types clearly. Although figure 6.16 does show the types of data and relationships more clearly than figure 6.13, the large edge in the centre of the diagram obscures some of the other edges. To overcome this problem figures 6.17 and 6.18 were created.

In figure 6.17 each edge has its weight mapped to edge width, but the ignore outliers function has been enabled. This produces a very different visualisation as the large edge has been scaled, allowing the weights of the other edges to be seen. The layout of nodes has been rearranged, so network 10.2 appears different in figures 6.16 and 6.17, but both represent the same data. Figure 6.17 illustrates three key relationships among several circular nodes with large dark edges. The circular nodes represent item types

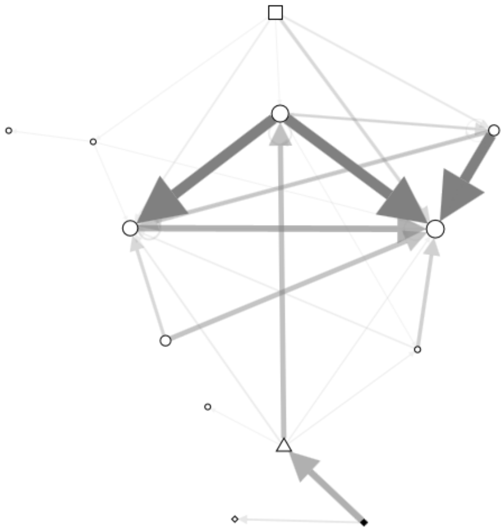


Figure 6.17: Network 10.2 rearranged

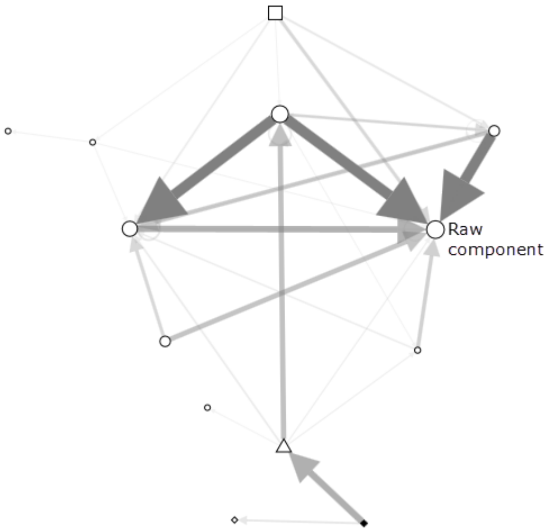


Figure 6.18: Network 10.2 with raw components labelled

and are connected by the largest, darkest edges, indicating the highest number of links, and arguably most significant relationships in the network, are between CAD items. There is also one node which is visually striking in the middle right of the diagram. For simplicity this has been labelled in figure 6.18 and is the node representing 'raw components'. Even without the labels present in figure 6.13 it is clear that this node is the most influenced in the network, with many edges travelling to it. In a company's data structure it would be expected that the most important links and the most influenced data would be part of the CAD collection and it appears that network 10.2 in figures 6.17 and 6.18 confirm this. A problem with figures 6.17 and 6.18 however, is that they somewhat misrepresent the true weight of these links. Therefore figure 6.19 was created.

In figure 6.19 network 10.2 is laid out in a very similar way to figures 6.17 and 6.18, however the edges now have colour and the opacity varies. This is due to changed visualisation of the metrics. While the edge width has been kept the same (decided by the weight of the edge with 'ignore outliers' enabled) the opacity is determined only by the edge weight. This allows the information about edge weight to be conveyed more correctly. While the width mapping allows all edges to be seen, the opacity highlights the true values (not scaled by the 'ignore outliers' function) producing a visualisation where the largest edge can be distinguished from the others, without causing them to be obscured, as in figure 6.16. The colour of the edge was determined by the type of relationship it modelled. In this multimodal network there are 5 different relationships, visualised simply in figure 6.11, but for this visualisation each was assigned a value

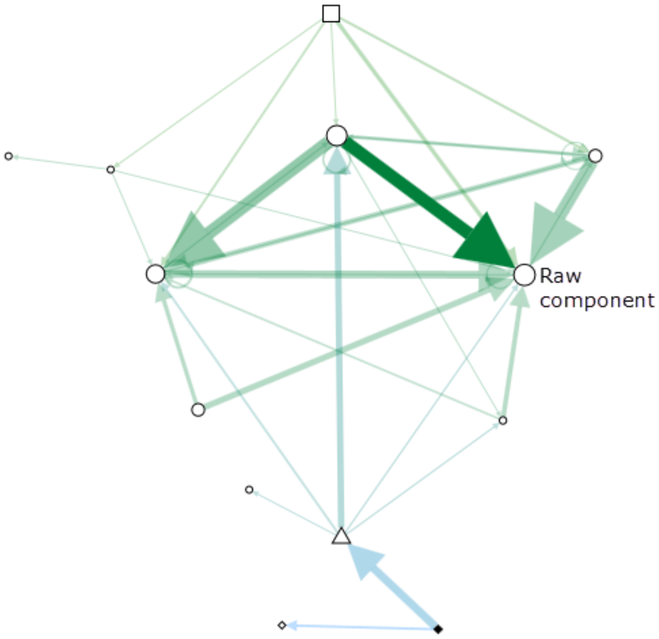


Figure 6.19: Network 10.2 with edge colour applied

which was used to determine the edge's colour on a blue-green spectrum. Again, while not definitely showing the type of relationship the edge represents, this colour difference indicates to the viewer that not all edges represent the same type of link. Throughout this thesis visual properties have been used to enhance network diagrams and figure 6.19 appears to indicate several of them are useful for communicating differences between the data.

This final mapping of network 10.2, shown in figure 6.20 has been edited so several metrics are mapped to visual properties, many of which have been varied from the previous diagrams. While the network in figure 6.20 still represents the same data as the previous diagrams of network 10.2 (figures 6.13-6.19), this diagram illustrates the need for a key or explanation when many visual properties are determined by metric assignment. While the interpretation of this diagram may be clear in this context, given the chronological discussion of this investigation, if it were presented without any further information, it might not communicate any information clearly to the viewer. In this way figure 6.20 shows that advanced network diagrams may require additional information in order to be useful and correctly interpreted, but indicates that network visualisations are powerful tools for communicating CAD data.

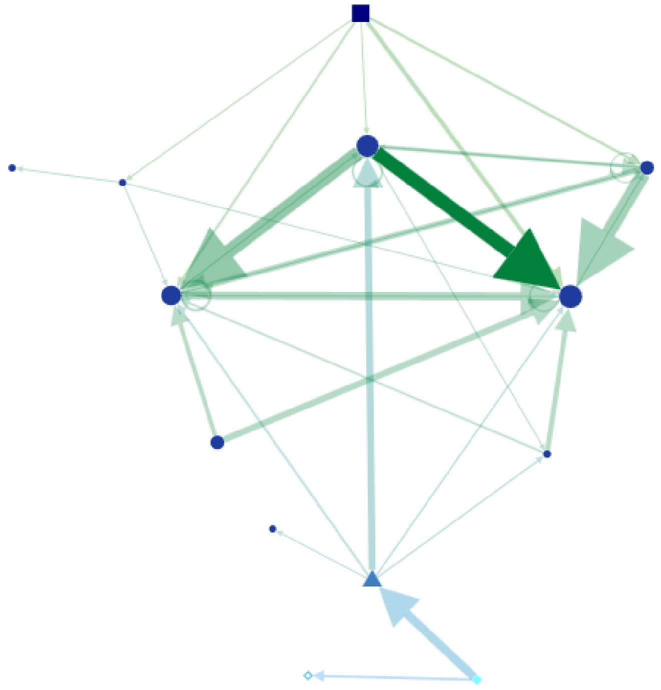


Figure 6.20: Network 10.2 mapped with many metrics

6.4.4 Company B data structure discussion

Using network theory to create a network exploring the structure of company B's CAD collection and associated data was an effective way to reveal the composition of the data. Network 10.1, presented in figures 6.11 and 6.12, clearly illustrated how the CAD data was related to the orders the company received and the customers who received the products, as well as the purchases company B made. Expanding network 10.1 to create 10.2, presented in figures 6.13-6.15, showed how the types of CAD data fitted into the structure of the whole data. As well as revealing further and previously unknown details about the structure of the data. Notably the global metrics were greatly affected; the network diameter increased but the density remained similar. Network 10.2 included more detail about the structure of company B's CAD collection and how it was directly related to costs via purchases and orders. Further exploring the visualisation of network 10.2 revealed that metrics could be useful when applied to visual properties, but that using too many may confuse the diagram and mean it is not able to be interpreted clearly by the viewer.

Comparing network 9.1, from figure 6.3, and network 10.1, from figure 6.11, similarities can be seen. Both contain nodes for types 'item' and 'order_line' with a directed edge travelling from 'order_line' to 'item' labelled 'order_line_is_for_item'. This suggests that it is a common relationship found as additional data to a CAD collection. The node labelled ‘source’ from network 9.1 is not found in network 10.1, and this makes sense because this node refers to specific data involved in project A. The nodes not found in network 9.1, but visible in network 10, include ‘customer’ and ‘order_line’.

Comparing network 10.1 to the predicted network shown in section 6.2, figure 6.1 shows further similarities. Nodes representing the CAD files, orders, purchases and customers appear in both, which are named differently. It is notable that in both the largest node represents CAD data, suggesting that in industrial data collections CAD data is the most important and most influenced part of the network. In network 10.1 it can be seen that the order node is connected by two out-edges to customer and order line, suggesting that for company B orders are the most influential part of the network. Network 10.1 does not include nodes of type ‘designer’ or ‘responsible engineer’, however this data was not provided explicitly by company B nor was it extracted from metadata in the CAD files on the grounds of preserving anonymity.

Illustrating this type of data with a network is a novel concept and the diagrams presented in this section are original representations of real world CAD collection data. Mapping the different metrics to visual properties or recording the most influential nodes within the network could give company B knowledge about their data they would otherwise not be able to access and indicate where efforts could focus for optimisation. Mapping the purchase orders to types of item, as shown in figure 6.14, could show company B where to focus efforts in order to reduce costs, while adding cost data to each node type could be a way to map cost flow. If metrics could be associated with actual monetary values, these network diagrams could illustrate the spread of finance according to actual data structure.

6.5 Data Visualisation

Data visualisation has becoming increasingly important and it is vital that the mechanical engineering industry and education sector take note of this development, and fully exploit the advantages it provides. McCandless commented on the importance of data visualisation during his 2010 TED talk, stating “By visualising information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you’re lost in information, an information map is kind of useful” [127]. The proliferation of data has meant a change in the way people interact with it and while advertising is an example of data visualisation, the attempt made to communicate large quantities of information clearly have resulted in more data visualisation and have inspired elements such as infographics, which are combinations of graphics and text.

Throughout this work original images have been produced, and several have already been presented and discussed, but many have not and this section will present those deemed notable and discuss their meaning. These images are visually striking, but are also important for communicating a wealth of information in an easily digestible way. Presenting them to someone outside of the field shows they can easily understand concepts. This is a key aim of data visualisation and infographics and as such, it is advisable that industrial companies or educators make use of such images to communicate with their audience. The images produced here may enable designers and engineers to interact with their own data in a new way. The vast quantities of data created by companies could be helpfully visualised using networks.

The diagrams produced throughout this research have illustrated network structure in various ways. Some basic visualisations were presented (e.g.: figure 4.1), while others made use of various layout techniques to present a clear visualisation (e.g.: figures 4.8 and 4.9). In many different ways metrics were mapped to visual properties to enhance network diagrams (e.g.: figures 4.6 and 4.12) and allowed novel interpretation and unique results to be explored (e.g.: section 5.3). In section 4.9 the various layouts available in NodeXL were discussed and examples were shown (figures 4.13 - 4.18). The various visualisations produced throughout this work have been both visually striking and provided unique insights into the modelled data. In previous chapters images have been used as illustration and examples of presenting data. In these preceding chapters these images have been solely utilised to demonstrate and convey the work and results that have been found as this research has progressed. This section will further discuss some complex network visualisations as these images are not only important for the purpose of showing the performed analysis, but are also important sources of information in their own right. Each image represents a set of data and can be used to explain and illustrate different things about the collection of data it represents.

During the research conducted on industry data, visualisation was further explored. The diagrams presented in subsection 6.4.3 discussed the merits of enhancing a diagram with metrics determining visual properties, presenting different images (figures 6.16-6.20). These were determined to be useful for strengthening the representation of data and could be seen to improve data communication. In continuing this investigation, the large network 9.2 was used to explore several layout options. Figures 6.21 - 6.25 show five of the most successful diagrams produced. In all cases network 9.2 is a multimodal directed network, the edges are highly curved and bundled and their width is determined by associated weight and nodes are sized according to their in-degree. In each diagram the nodes have been coloured, but the colours were chosen to enhance visualisation instead of signifying any particular characteristic of the data.

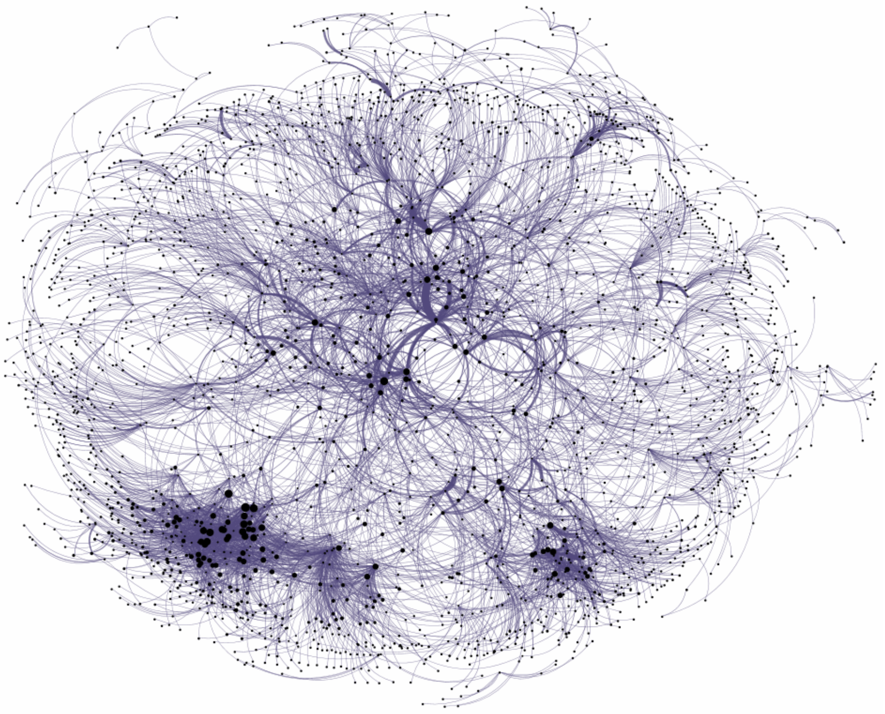


Figure 6.21: Network 9.2 in Fruchterman-Reingold layout

Figure 6.21 shows network 9.2 in the Fruchterman-Reingold layout, with the edges coloured purple. In this layout there are some nodes which appear clustered together while others are on the outer edges of the visualisation. This is likely due to the algorithms used in laying out the nodes and could indicate that they are more highly connected than those which appear sparse and distant from the whole.

Figure 6.22 shows network 9.2 in a circular layout. The bundled edges have been coloured orange and result in a space in the centre of the circle, unlike the circular layout shown in figure 4.10. The larger edge widths are clearer than in the Fruchterman-Reingold layout shown in figure 6.21, but the random layout of the nodes may not indicate those that are more closely linked than others. Also as a network becomes more dense it becomes increasingly challenging to see individual edges linking individual nodes. This is evident by comparing figure 6.22 to figure 4.10.

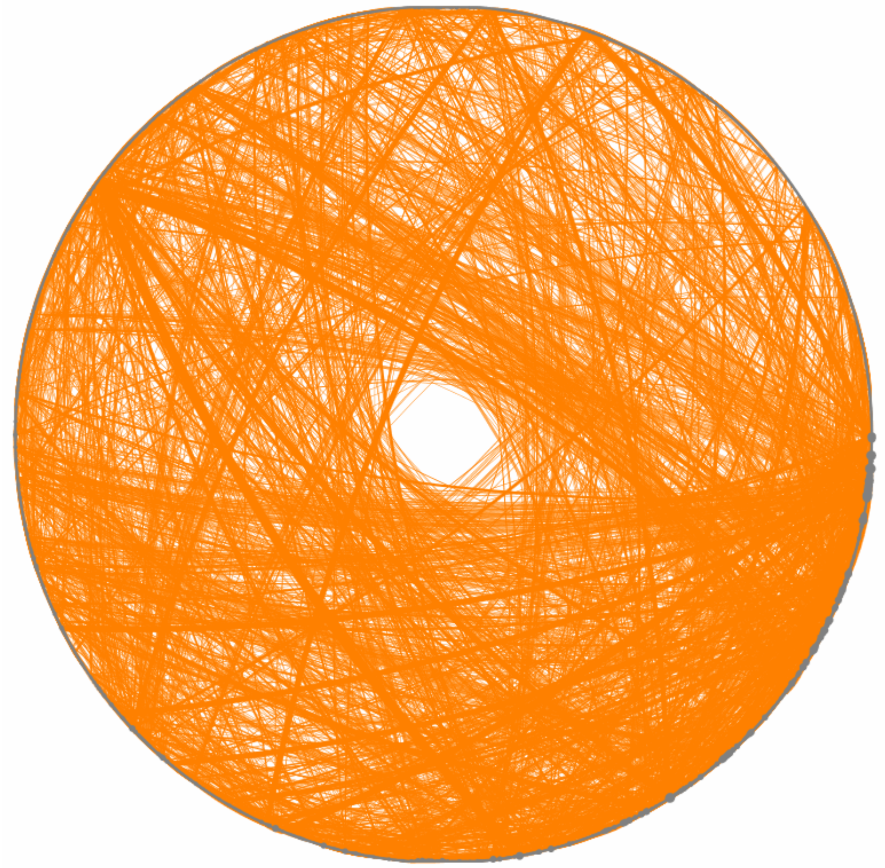


Figure 6.22: Network 9.2 in circular layout

Figure 6.23 shows network 9.2 in a grid layout, with the edges coloured purple. In this layout the nodes are placed in a grid formation at random, they are not placed next to those they have stronger links to, as can be seen from the spread of the nodes. Again large, heavy edges can be seen linking nodes, indicating strong relationships, alongside lighter, thinner edges, showing lower value relationships. In this layout some of the large nodes are clearly visable, especially when compared to the circular layout of figure 6.22, but unlike figure 6.21 the nodes are not grouped making the image potentially less useful. The grid-like structure may not provide a characteristic shape for a company's data however the spread of edges between nodes and the depth provided by their opacity might make it more possible to compare different networks to each other. These could include networks made from individual company's data collections or different stages of a single company's database, to allow comparison and characterisation of a CAD collection over time. This could provide a novel way for large collections of CAD and associated data to be overviewed and compared.

Figure 6.24 shows network 9.2 in a Harel-Koren Fast Multiscale layout and the

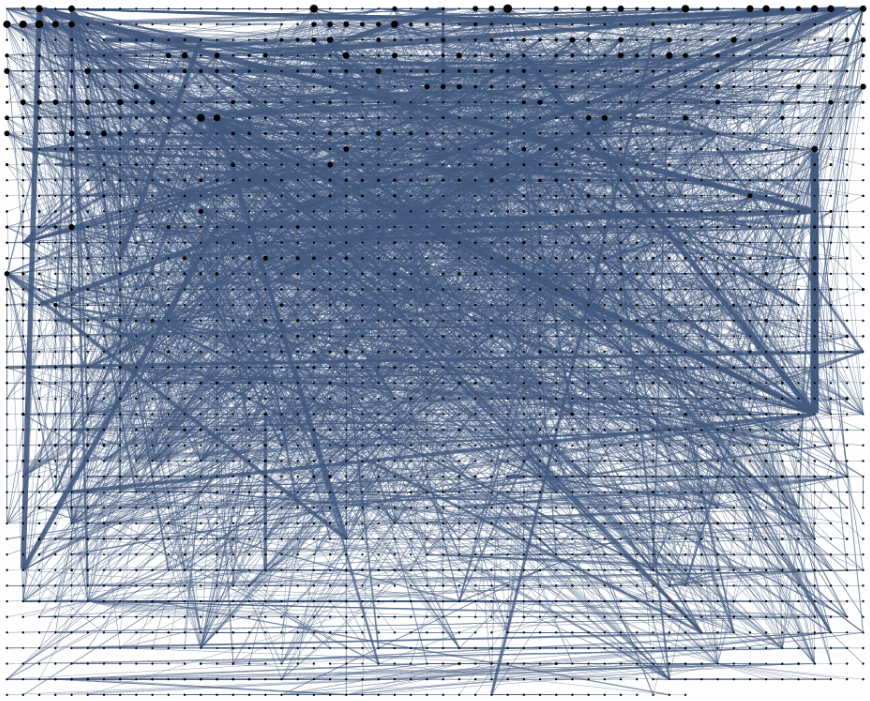


Figure 6.23: Network 9.2 in grid layout

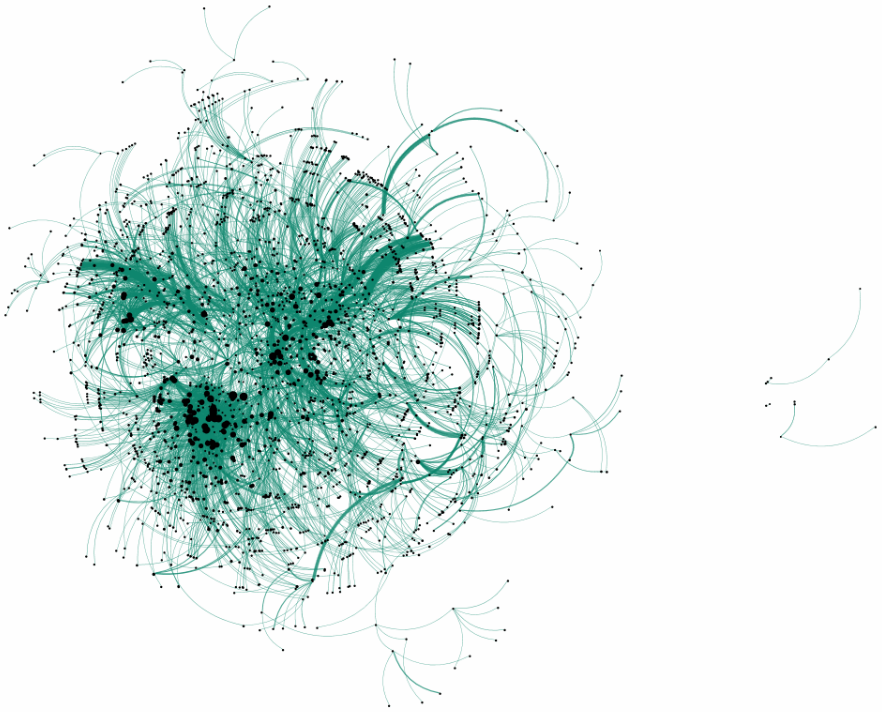


Figure 6.24: Network 9.2 in Harel-Koren Fast Multiscale layout

edges are coloured teal. The first striking thing about this layout is how four separate components are visible, one to the left of the image and three much smaller ones on the right-hand side. This large collection on the left clearly bunches several nodes together, again suggesting the importance of their relationship. The three separate elements on the right-hand side of the diagram highlight the partial nature of network 9.2 (which was created from 15000 randomly selected lines from the graph database of company A's data). They do not indicate that there are some unconnected parts of the data, only that this is a partial model. Notably this is the only layout to show this clearly, suggesting that if a company wished to check for disconnected data elements in their system, the Harel-Koren Fast Multiscale layout would be an effective and quick way to visualise this.

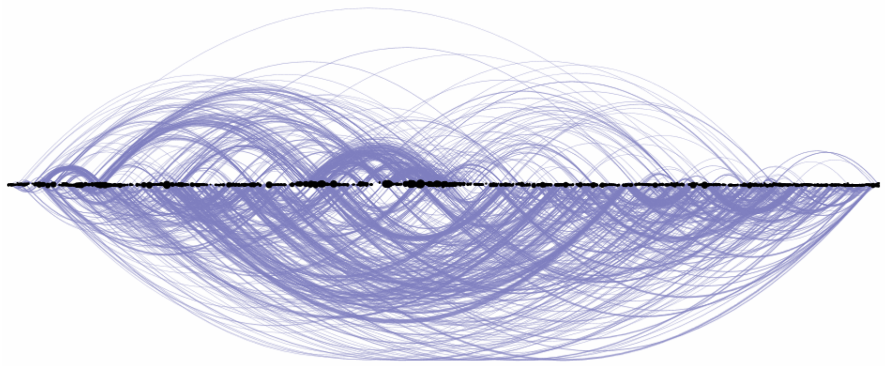


Figure 6.25: Network 9.2 in Sugiyama layout

Figure 6.24 shows network 9.2 in the Sugiyama layout, known as an ordered tree layout, with edges coloured purple. The resultant image is unexpected; the nodes are ordered in a line, rather than a tree or layered structure, akin to that shown in other examples of Sugiyama layout, such as figure 4.3. Despite the ordered nature of this data, this is not communicated effectively by this layout. The nodes are arranged in a straight line and without the curved edges would appear to be connected by a single line. The layout provides an interesting visualisation, but does not communicate data about the nodes well. The edges are the most striking part of this visualisation and the heavy, large edges clearly. This layout may be successful in communicating information about the types of relationships edges represent in a multimodal network, especially if a colour code was used.

Comparing the five different layouts presented in figures 6.21-6.25 shows the importance of carefully visualising the data structure that a network represents, as different layouts can give different impressions and highlight different characteristics of a network. Figures 6.21-6.25 suggest that the grid layout may be most effective for comparisons between CAD and other associated data collections, while the circle layout may show little meaningful data as a network becomes more dense. The Harel-Koren Fast Multiscale layout may be the most effective layout for discovering visually if a large network has unconnected elements, however this can also be shown by calculated metrics. The Fruchterman-Reingold layout could be a successful visualisation for characterising a data collection as well as identifying clusters of highly linked nodes within a network.

During the course of this research, relevant network visualisations produced were presented to the company A for their consideration. The network diagrams were used as an illustration, aiding ShapeSpace in their customer engagement, and were included in progress reports while findings were presented. The diagrams were well received and companies were receptive to them. Illustrating company data structure in this way is an original concept, and despite the capabilities of the PLM systems companies employ, they are often unable to see their data in this way. These visualisations suggest that network diagrams displaying metrics visually may provide a way to characterise company data, and would reveal new insight to companies about their own data. While manufacturers are very protective of their data, the rapid production of visualisations that have a high impact and convey information effectively could allow these diagrams to quickly indicate the necessity of this analysis to companies.

Throughout this thesis layouts were chosen based on clarity. While network diagrams clearly represent a vast quantity of information, they are also visually stunning and provide unique and hitherto unseen visualisation of company data. However investigation has shown these diagrams provide much more than visually interesting images. The powerful communication and analysis these figures allow indicate they could be of high worth, especially when presented alongside calculated metrics.

The value of data visualisation was discussed in section 2.4 and therefore it is suggested that these network visualisations be considered for further investigation. Throughout industry it is desirable to clearly communicate information internally and publicly. Whilst ShapeSpace used these images to engage their customer it would be possible for company A to use these images to communicate information clearly within the company and to their customers. Production of these network visualisations is rapid, indicating that they could be high value and low cost analysis methods. An interesting aspect of these network diagrams is that it is possible to interpret them clearly without advanced knowledge of the data they represent. They might be used effectively in communicating a lot of information, particularly to the public with little prior or specialist knowledge.

6.6 Discussion

In this chapter, original work investigating the use of networks in modelling and assessing real world company CAD and associated data is displayed and results presented. Mechanical engineering design and manufacturers originate some of the most brilliant and creative design, using and producing ground breaking technology, methods and products. The use of network theory could have a positive impact upon the industry if this work is continued. It is suggested that calculated metrics and visualisations could be effective in categorising and comparing CAD collections and entire data structures, while the use of networks to underpin graph databases allows further analysis.

When assessing company A's data, it was suggested that network metrics could be used as a self-checking function. The network built to model company A's entire data structure originally contained a bug that resulted in an incorrect model. However the calculated metrics quickly revealed there was an error, discussed in subsection 6.3.3. As the network was rebuilt, the progression of the investigation was presented in the rest of section 6.3. It could be possible, if analyses like this were to be developed, that network metrics could be used to check the networks modelling CAD collections had been built correctly, before further analysis.

Significant models of a partial and whole scope of company A's data are presented and discussed in sections 6.3.5 and 6.3.6, shown by networks 9.2 (figure 6.9) and 9.3 (figure 6.10). Network 9.2 investigated a large partial network model of the data, whole network 9.3 went on to advance this to a complete model. This large model showed a network of a CAD collection with associated data had a smaller diameter and a smaller density than a network of the CAD collection alone, which may have implications for search methods and other analysis capabilities.

This exploration continued and data from company B was modelled in similar ways. This data included many more types than the previous analysis and various visualisations were presented and discussed. It was suggested that while metrics mapped to visual properties could provide additional information about the modelled data, it was key that this be clearly explained or limited as too many additions made the visualisation overcomplicated.

In section 6.5 five original diagrams are presented of network 9.2. These network visualisations show a large collection of company data, mapped using different layout options. The various layouts shown highlight different aspects of the structure of the company data, indicating that further work could continue to explore the powerful capabilities of network diagrams in communicating data clearly even outside industrial contexts.

Networks of CAD collections, such as network 5, could be instrumental in improving design reuse methods. Using network methods to create a queryable database, as was used in section 6.3 to explore company A's data structure, could provide new capabilities for design reuse methods. While not explored here it can be seen that a network may provide a suitable basis from which to pilot a new design reuse technique. Network 9.2 suggested that expanding a network with additional data structure would increase the diameter of a network, though not its density. How it was seen in network 9.3 that this additional data structure decreased the diameter of the network, agreeing with network theories that suggest including additional nodes and edges in a network aids traversal. Therefore it could be beneficial to explore the uses of whole data structure models with relation to CAD search.

6.6.1 Critique of recent similar work

Work done by Mill [14, 47, 131, 132] has explored industrial CAD collections and the Edinburgh Benchmark collection. Bimodal and multimodal networks have been built, with an emphasis on grouping CAD files by feature, providing different links for network analysis. This work emphasises an alternative data structure from the work here, using metadata from the CAD files to create a differently detailed network. The networks effectively model CAD, showing how shared feature use can link different designs. Using features to categorise parts is not uncommon in design work, especially as many CAD models contain a 'feature tree'. It is suggested this method could be combined with the multimodal networks presented here to create a powerful network model of a CAD collection.

6.7 Conclusions

The aim of this work was to explore the uses of network theory when applied to collections of 3D CAD model:

> Produce a network of a real world parts collection from the manufacturing industry. From this, collect information about the uses of network theory in

this context. (From aims and objectives in section 1.3)

and that has been achieved by building networks of real world industrial data from two companies. Furthering the work presented in chapter 4, this chapter has investigated these data collections, exploring basic and more advanced network models.

Each data collection was modelled by a simple overview network, shown in figures 6.3 and 6.11. These were expanded and the first study focused on developing a network where individual CAD files and orders were represented by nodes, while the latter prioritised showing the 'types' of CAD models according to the company's data structure.

The original networks presented here reveal several potential uses for network analysis in industry. Network theory provides a novel measure for the size of a company's CAD collection and associated data, via the metrics of diameter and density. While ShapeSpace have piloted using network databases to assess clients data, these statistics have not been applied to data held within PLM systems before and might be useful in categorising CAD collections. It is suggested that these measures could be key in advancing search techniques, particularly the issues surrounding searchability of a CAD collection.

This research has investigated new methods for exploring and visualising real world company data. Suggestions have been made as to the use of these methods, while network diagrams representing company data structure have been presented and discussed. These real world data structures show how a CAD collection is linked to other company data, highlighting that orders are can be considered the most influential data within a company, while CAD files are the most influenced. This might have been expected based on knowledge of manufacturing industry, but these results effectively verify the assumption.

In mechanical engineering design industry, CAD is highly valued data and the rendered images produced are often cutting edge. The visualisations presented in this thesis could be considered akin to the advanced 3D models these companies pride themselves on and the worth of this analysis is increased by the high value data represented. In this context network diagrams provide an innovative and unique way to communicate CAD and associated data structures and information, providing a network view of manufacturing environments in an original way.

6.8 Further Recommendations

This work has explored the uses of network theory to model collections of CAD data from industry, presented novel methods and visualisations and suggested several areas for continued investigation.

Further work on large collections of CAD models could seek to assess the whole database belonging to a company. Network 9.3 is a multimodal network of CAD and other company data, but it would be possible to include other industrial data, such as customers, purchases and designers, to create a more complete network of a company's data. This could be a large multimodal network, although it would be computationally expensive. The metrics calculated could be compared to those for a CAD collection network and a CAD collection with additional data, such as networks 5 and 9.3, to assess the influence of additional company data on a network model. Also other real world data could be used to build networks similar to those presented here, or to this suggested next network, to allow comparison between company CAD collections. This could allow characterisation metrics to be identified.

The graph database used in section 6.3 and the associated network provided a useful basis on which analysis and queries could be run. It is suggested that work could continue with this model, using TinkerPop, Gremlin and other associated software to explore company data structures. This method allowed errors to be found, partially due to a bug in the code, but also because real world data is not as expected. Often standards are not conformed to and data is incomplete and other analysis happens within a characterised context. This method could be helpful in detecting different kinds of issue, without the necessity of defining the manufacturing environment.

It is also suggested that the visualisation work begun here should be continued. Infographics are a popular and lucrative data presentation technique and the work here presented here shows the value of the information rich images resulting from network theory. These images could be effective at communicating companies' designs and ideas internally or to the public. The tools for visualising networks could also be further explored, so layout options could be better understood and comparisons made between companies' data or between a single company's data structure at different time periods.

Chapter 7

Conclusions

CAD collections can be effectively modelled and assessed using network theory; the methods presented in this research reveal new information about CAD data structures and the associated metrics allow further analysis not previously applied to mechanical engineering. The ten original networks presented in this thesis include a range of CAD collections, modelled as nodes and edges, and presented with their associated diagrams and metrics. CAD data has not been presented in this way prior to this work or outside of Mill's research group and doing so provides new capabilities for mechanical engineering design.

An overarching aim was stated at the beginning of this thesis and through initial investigation of CAD networks several areas for additional study were proposed, as presented in chapter 4. This lead to an exploration of unimodal networks for modelling CAD from education and similarly of multimodal networks for representing industrial CAD collections, in chapters 5 and 6 respectively. The other areas identified for further work which were not pursued here are discussed briefly in section 7.2.

The first of the presented networks was the simplest and smallest; network 1 showed one CAD assembly, where the CAD files are indicated by nodes and the edges represent the 'contains' relationship between the models. From there each network presented was more complex or larger. Network 2 modelled two mechanical assemblies that shared simple components, while network 3.1 was of the Edinburgh Benchmark collection. Comparing these networks showed that when CAD data is modelled by assembly structure a medium sized collection of files may not be connected as one network. Larger networks, such as 5 and 9.3, revealed that industrial CAD collections are connected as one element. This suggests that industrial CAD collections are well connected, even when modelled by assembly structure.

After modelling CAD files linked by 'contains' relationships in their assemblies, shape similarity assessment results were used to build network 4, and so produce a network where the relationships modelled were 'is geometrically similar to'. The data used for this investigation was taken from an educational setting and so the nodes were simplified to represent a 'folder' belonging to one student, rather than individual CAD files. The results from this initial network were significant, especially when considering plagiarism, therefore further networks were built, as presented in chapter 5. The novel method presented allowed students' CAD to be assessed for geometrical similarity, which was deemed to be an accurate measure of plagiarism in 3D designs. This analysis showed that the metrics of betweenness centrality, clustering coefficient and heavy edges indicate unoriginal work. The proposed method was tested on historic student data from the University of Edinburgh and on a current class. These results suggest this new method is robust and reliable for highlighting suspicious 3D student files, which should be checked more closely by the human marker for plagiarism. There is currently no accurate or accepted way to do this for non-text work, though efforts have been made to identify unoriginal material in the visual arts.

Other investigations presented networks of real world industrial CAD data. Network 5 models the whole of a company's CAD collection, while the largest network presented, network 9.3, models that CAD collection and adds all its associated orders. These original networks are notable for their sizes; network 5 contains 13125 nodes and 65007 edges and network 9.3 has 17306 nodes and 201840 edges. The use of data from engineering manufacturers allowed the designs to be assessed in an entirely novel way, and showed that the diameter and density provided innovative measures of a CAD collection.

Unimodal, bimodal and multimodal networks have been presented throughout this investigation. For simplicity when this investigation was begun, the networks were viewed as unimodal, allowing all global metrics to be used. This is reasonable when looking at CAD files, because assemblies, sub-assemblies and components linked by 'contains' relationships can be considered similar. In small networks this is acceptable. However to better understand large collections of mechanical engineering designs, it may be pertinent to consider these networks as bimodal or multimodal, depending on the data modelled, as the influence and importance of an assembly and a component part in a large collection may vary drastically.

This chapter will now evaluate whether the aim of this research has been met based on the set objectives, and go on to discuss the limitations of this work and make recommendations for further research.

7.1 Review of Aims and Summary of Thesis

In section 1.3, the aim of this thesis was stated:

> The central aim of this work is to explore the uses of network theory when applied to collections of 3D CAD models and its uses when combined with shape similarity techniques, in a mechanical engineering context. This will be supported by investigating two main area where CAD models are found, namely in manufacturing and education, both key areas of mechanical engineering.

and this has resulted in three related investigations; an exploration of network theory applied to various CAD collections, an in-depth unimodal model of CAD in an educational setting and an analysis of multimodal networks of industrial CAD collections and associated data. The central aim of this work has reportedly been achieved in several ways. In chapter 4 five original networks were presented, along with diagrams and metrics which were analysed. This fulfilled the first objective of this investigation, to

> Produce networks of parts collections from different scenarios, using shape similarity and other linking techniques to develop networks of CAD models. Use network metrics to measure the created networks and identify from network theory the most useful metrics, for use with CAD part networks.

Several interesting results were shown and illustrated that networks provide an innovative way of mapping and measuring CAD collections of different sizes. It was found that CAD networks displaying ‘contains’ relationships do not obviously conform to Milgram’s ‘six degrees of separation’ theory nor Watt’s small world theory. It was demonstrated that metrics could be used to highlight novel characteristics of CAD collections; in particular the diameter and density of a CAD network.

Following these discoveries, it was suggested that further investigations of CAD in education and industry would worthwhile. Chapter 4 examined unimodal networks of students’ CAD files and proposed a definition of plagiarism in CAD education. These networks (network 6.1, 6.2, 7, and 8) were built using ShapeSpace's geometric similarity technology and therefore were entirely novel. This investigation fulfilled the last objective of this work, which was to

> Develop the uses of networks in an educational setting, by investigating a class's design submissions, focussing on the role of networks and shape similarity techniques to assess the similarity of students' work, and thus develop a way to identify plagiarism within a class's submissions.

and reported that it was possible to identify plagiarism using network theory. It was found that the metrics which highlight possible plagiarism are large betweenness centrality, low clustering coefficient and the presence of heavy edges connecting students, however, as with all detection techniques, human consideration is needed to finally confirm the presence of unoriginal work. A real world data set from a current class at the University of Edinburgh was used to test this conclusion, and the marker reported that the result of the network analysis assisted the grading process. The discussion in section 5.7 considered that these results could effectively locate copied CAD files, however it is decidedly unclear whether they will help confronting the issue of cheating within the educational system. It was concluded that this method will, however, aid educators in providing marking with academic integrity.

Turning to industrial CAD collections, chapter 6 explored the uses of network theory when modelling real world CAD parts and associated data. These multimodal networks and their diagrams are entirely novel, fulfilling the third objective of this work:

> Produce a network of a real world parts collection from the manufacturing industry. From this, collect information about the uses of network theory in this context.

It was found that networks provide interesting analytic possibilities, and can be used to ensure that databases are correctly built. Industrial data structures were shown to be well represented by these graphs and the metrics of diameter and density are proposed as valuable new measures of a company's data collection. An extensive model of an industrial CAD collection with associated order information is presented and, due to the metrics of this network, it is suggested that large data models could assist search and design reuse methods. Visualisation is also discussed, with several original network diagrams presented, demonstrating the advantages and drawbacks of different layout options when dealing with large data collections.

Throughout this work novel images of network diagrams are presented and associated metrics discussed. In section 4.3 a definition for these measures, when calculated from a CAD collection, is suggested and the investigation in this thesis found these to be effective explanations. Visualisation of network diagrams is discussed throughout this thesis, identifying the key benefits of these images, including the efficacy of mapping metrics to visual properties in order to clearly communicate the data. This goes some way to fulfilling the second objective of

> Classify the use of network theory for analysis, search and visualisation of the created networks.

however this research does not include sufficient investigation to prove the usefulness of network theory techniques in defining the searchability of CAD collections. Therefore this, among other steps, is suggested for further investigation.

7.2 Limitations and Scope for Further Research

This thesis has several limitations and there are various areas where further work is suggested based on the findings presented. It is noted that one basic objective that is not accomplished, namely the verification of network techniques assisting search methods for CAD. Despite this, it is considered that the research presented here is an in-depth investigation into network theory when used to assess CAD collections. This is achieved using the Edinburgh Benchmark set, as well as real world data from education and industry. It is evident throughout this thesis that network theory provides many novel methods, measures and approaches when analysing CAD collections and this research is considered successful in achieving its aim.

Chapter 5 demonstrated that metrics can highlight possible cases of plagiarism in student submissions on CAD courses. This was demonstrated through simulations and a real world test case. However this method is involved and laborious, requiring specialist knowledge to produce the results. Therefore it is recommended that these discoveries be considered for inclusion in plagiarism detection software. As several key metrics were found to unveil potentially suspicious work, these could be used as the foundation for building a software package, which could take a collection of CAD submissions and, using the method documented, produce the results for a user. This software could be built with a simple interface, allowing widespread use without requiring an understanding of shape similarity or network theory, and could be made available to educators. A suggested roadmap for this is to use ShapeSpace's legacy tools, adding this functionality to build a program which could then be tested by staff at universities on real world students' CAD collections. It is also possible that this type of software could be developed for use within industry, and considered for use in either cases of plagiarised designs that may occur online or in competitors' CAD collections, or for use by designers, so they could check for similar designs.

In chapter 6 it was suggested that networks could provide enhanced search, and therefore design reuse, capabilities to manufacturers. However this was not explored in depth and no classification of the use of network theory for search was made. It is suggested that this work continue, using diameter and density metrics as a starting point from which to calculate a value for the searchability of a network. While this investigation has not furthered the fields of shape search and design reuse, it has proved useful in industrial settings, where company data is analysed. For this reason it can be seen that continuing to explore network theory to analyse real engineering data could provide a new and reliable way to effectively implement design reuse within industry. It is notable that Watt's reported approach of adding strategic nodes to a network can reduce the diameter significantly [194, 197]. It could therefore be pertinent to explore this claim in relation to search in a network.

In considering the uses of network theory in industry, a potentially very promising line of investigation would involve adding cost data and profit statistics to node properties. It is thought that including financial data in network analysis may provide valuable measures and quantitative information to companies and may assist in cost cutting efforts.

Comparing network 3.1, Edinburgh Benchmark parts linked by 'contains' relationships to network 5 and 9.3, networks of real world company parts also modelled as connected by assembly structure highlights one clear difference. Network 3.1 appears as 24 separate elements, but the real world data sets form one connected element. This may be mainly due to the sizes of the collections; the Edinburgh Benchmark contains 245 parts while the data sets used for network 5 and 9.3 contain 13125 and 17306 respectively. While considering networks 5 and 9.3 it is vital to remember the CAD data contained in both is the same, but network 9.3 includes additional associated industrial data, so is not directly comparable to network 3.1. However, it is notable that the large collection of CAD files from industry is connected, unlike the Edinburgh Benchmark. These differences suggest that there is need for an additional benchmark to be created, modelling an industrial collection of designs. This could be generated from the Edinburgh Benchmark or from original parts. It is suggested that this benchmark be tested with network analysis to assure it accurately represents industrial models. This would allow further research, using network theory and other methods, to use a robust benchmark for modelling a real world engineering manufacturer collection.

Throughout this thesis the many novel visualisations of networks have been arranged to clearly illustrate the modelled data. These images have proved effective and have been presented at conferences and to one of the customers whose data was modelled. The positive feedback received suggests that this work could be expanded upon. While network diagrams are not uncommon, they have not been produced to represent CAD data outside of Mill's research group. They are considered unique contributions to the field and it is suggested that their uses as effective communication tools be explored. They could be most effective in engaging companies who wish to analyse their data and CAD collections, and used to identify the different issues and successes within a large database.

This research has not considered some advanced aspects of network science. Several examples of these are scale free networks, the spread of ideas, preferential attachment, and network robustness. It is suggested these be investigated to develop the effects this powerful theory provided in assessing CAD collections. While many of the networks presented here may be considered scale free, this has not been explicitly discussed. It is advisable to consider the structure of the large CAD networks further, as several might be scale free networks. Investigating this could lead to the discovery of important hub nodes, which connect much of the network, and this could have a large impact on search capabilities. Using network analysis to locate these hub CAD files could allow new insight for search and retrieval techniques.

Fluid networks are another area that has not been considered in this thesis. In a network it is thought that the data represented by nodes is not independent, due to the nature of the data structure. This is observably true in a social network of friends where one person's actions impact another, or in biological systems where reactions of one cell affect another it is linked to, but this effect is less plainly observable in a network of CAD data. In any real world situation, active part collection files will be edited and changed, however this analysis has not taken those changes into account. It could be possible to investigate the automation of network construction, so when local changes were made to CAD files, the network containing them would automatically reflect the alterations. This may have applications in improving data management within industry.

Appendix A

Publications

[1] Mill, F., Sherlock, A., Pan, Q., and Anderson, E. Recognising 3D products and sourcing part documentation with scanned data. *Computers in Industry 64*, 9 (dec 2013), 1201–1208.

[2] Anderson, E., and Mill, F. Detection of design reuse in 3D CAD using network analysis. In *6th International Integrity and Plagiarism Conference* (Newcastle, 2014), Plagiarismadvice.org.

[3] Anderson, Esmé and Mill, Frank. Better connected: 3D CAD networks (Poster). Presented at *The University of Edinburgh's Engineering Graduate School Conference 2014* (Edinburgh, 2014).

[4] Anderson, Esmé and Mill, Frank. Seeing things differently (Poster). Presented at *Dealing with Data 2015* (Edinburgh, 2015).

[5] Mill, Frank and Anderson, Esmé and Sherlock, Andrew and Corney, Jonathan and Paterson, Duncan. Network theoretic depictions and metrics for collections of 3D CAD models. Submitted to *Computer-Aided Design*.

Computers in Industry 64 (2013) 1201–1208

Contents lists available at SciVerse ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Recognising 3D products and sourcing part documentation with scanned data

Frank Mill [a,*], Andrew Sherlock [b], Qi Pan [c], Esme Anderson [a]

[a] University of Edinburgh, School of Engineering, King's Buildings, Edinburgh EH9 3JL, UK
[b] ShapeSpace Ltd, UK
[c] University of Cambridge, UK

ARTICLE INFO

Article history:
Received 10 September 2012
Received in revised form 19 March 2013
Accepted 28 March 2013
Available online 29 May 2013

Keywords:
Search by shape
Networks
3D scanning
PLM
CAD
PDM

ABSTRACT

Searching databases of 3D models is a crucial yet difficult problem that has been studied by the academic community for a considerable time. A useful and robust method for finding engineering parts remains difficult however. Previous work typically describes finding the best match in a single search. Work described in this paper uses scanning techniques allied to shape similarity measures to produce a system that successfully allows search by browsing. We also describe some new shape descriptors and methods of identifying and dealing with chirality. The technique is evaluated in the context of the part search applications. The use of the techniques is applied to large (80,000 + parts) databases of real world engineering components in use in automotive and aerospace companies. The methods employed are applicable to a wide range of scenarios in engineering, as well as the arts, archaeology, medicine and commerce.

1. Introduction

The scenario where a user holds an object in front of a computer and asks 'what can you tell me about this?' is where the work described here starts. More specifically it is aimed primarily at engineers who need to find data related to a product on a company network or intranet in situations where exact part names or numbers are not easily to hand. To perform such a search, or in fact a wide variety of searches, it is common to first formulate a query and then analyse the results of the query action and then perform refining further search.

In the system described in this paper fast efficient scanning is combined with a novel search engine that is based on part geometries and this allows the user to find files related to a physical hand held part. A number of input methods have evolved to formulate the query model and different strategies tried for the subsequent search, the best combination being dependent on the specific application.

The work described in this paper relates to applications in a wide area of product management situations, e.g. part information retrieval for design re-use, maintenance, marketing or user support. Rather than aiming to increase the fidelity of 3D scanned models the aim of the work was to enable fast and accurate identification of part data already stored in local, intranet or even internet based file stores. This may be in the form of a wide number of possible representations such as those found in CAD repositories, PLM systems or catalogues and these may be difficult to navigate due to the lack of exact part data.

The background of the work is based on previous studies developing systems to characterise shape [1–3] for applications in part classification and search. These systems calculate many key parameters of parts such as their surface area to volume ratio or their aspect ratio and these in turn are used to group or cluster part collections so that they can be easily searched. This enables rapid part retrieval without the need for exact part names or numbers. Shape based searches are useful for simply finding parts but they may also aid part database management by identifying duplicates or multiple similar shapes or they can be used to assist re-use of existing designs [4–6]. In general they can be used where downstream (from design) users require 3D part representations, e.g. manufacturing, maintenance or non-engineering functions such as marketing and customer support.

Searches for part data are often performed to find 3D models, drawings or other associated documentation such as manuals, analysis results or manufacturing plans.

Shape based search allows parts to be found when only approximate ideas of a part's shape are known. Fig. 1 shows some typical search strategies that are in use in the system described in

[*] Corresponding author. Tel.: +44 131 337 4779.
E-mail address: Frank.Mill@ed.ac.uk (F. Mill).

0166-3615/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compind.2013.03.019

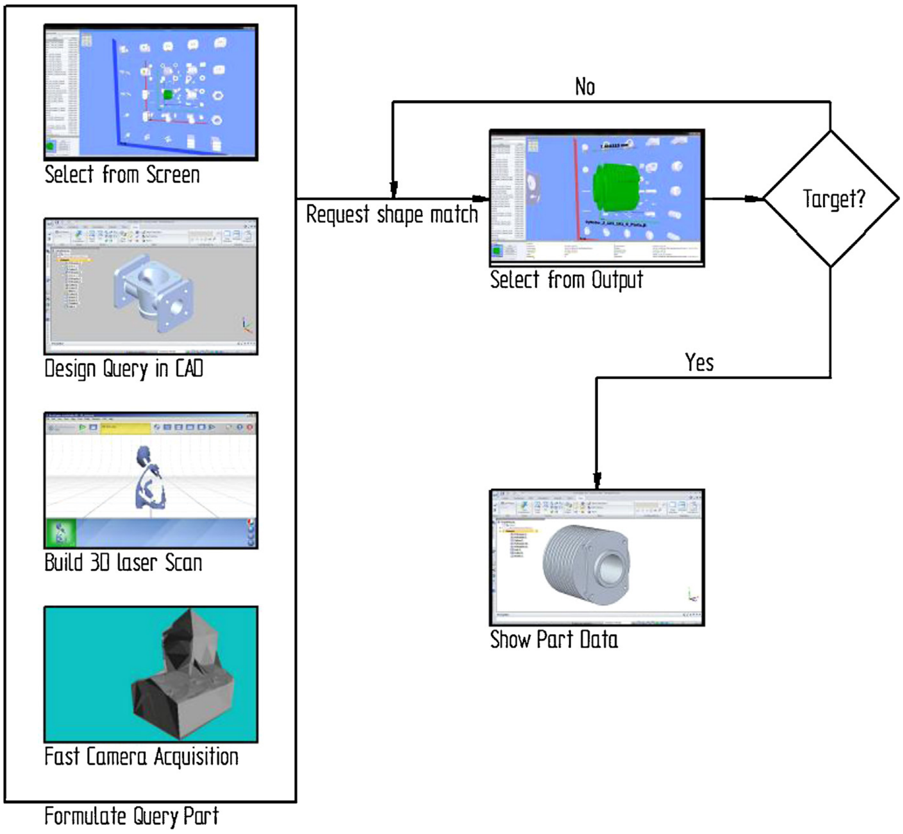


Fig. 1. Search flowchart showing how different query methods can be used to access a recursive search (browsing) process.

this paper. User searches can be carried out by picking from a bank of candidate on-screen shapes (select from screen) and subsequently selecting better matches from further screens of suggested examples or by designing individual 3D model queries (Design Query in CAD). Recently new methods have been developed whereby users employ rapid 3D scanning techniques (Build 3D Laser Scan and Fast Camera Acquisition) to build a query model from a physical part or part model.

Most research in 3D scanning use is aimed at producing high resolution scans and making high fidelity models from these, typical problems being interpreting actual geometry from the resultant point clouds. In the research presented here the emphasis is on fast scanning. The aim is to scan with just enough resolution to get a satisfactory query model and as a further development ordinary white light camera scans (e.g. from laptop webcams) have been used. The current system can search for part models and associated data that is stored in most 3D formats including STL, VRML, IGES, STEP and the majority of vendor specific formats for CAD and finite element analysis.

There are a great many applications for searching file systems and networks based on part shape, partly because the techniques allow access to data in a wide range of storage media including web or cloud based repositories.

In CAD environments where most of the users have considerable 3D modelling skills it often quickest to harness these and allow the user to design a query part from within his familiar CAD system. Even complex components can have very rough models constructed in seconds by experienced modellers. However, for users without CAD facilities, commonly the vast majority, this is not a favourable option and instead scanning methods may be employed. Although 3D laser scanned (and point probe) data have been successfully tried by the authors, the strategy suffers from the fact that in order to create useful scans specialised suitable equipment is needed and before the scan can be started is often necessary to spend time setting up a part. Research has been carried out by Pan et al. [7] that seeks instead to build 3D models direct from simple camera images that can be constructed using common devices such as those available on most computers and laptops. What is described in this paper is a system that optimises the general search process for a user. It is our view that much of the previous work described in the literature, e.g. that reviewed by Tangelder [2] works on the assumption that search will be performed on a general set of shapes. These are frequently general objects that are to be located and differentiated between (e.g. aeroplanes, cars, furniture and animals). This classification based view has seen the development of test part databases of general shapes which often exhibit shapes that are not relevant in manufacturing environments and which often have subsequent problems that are rare in the real world. It is rare for example to find a CAD designed engineering part that is not designed on a major *x*, *y* or *z* axis.

2. ProFORMA

The ProFORMA system takes solely a live video feed from a webcam as input, and contrary to many other reconstruction systems, aims to build a coarse 3D reconstruction for immediate use, rather than an accurate 3D reconstruction for later use. This makes it ideally suited to performing as the front-end for reconstructing a 3D model of a query object in a search environment, where it is desirable to get immediate results. Additionally, the ProFORMA system is designed to be used with the camera in a fixed position, with the query object being rotated in front of it. This has the natural benefit of being able to segment the object of interest from the background, something which is still an issue for systems which involve the inverse scenario of a moving

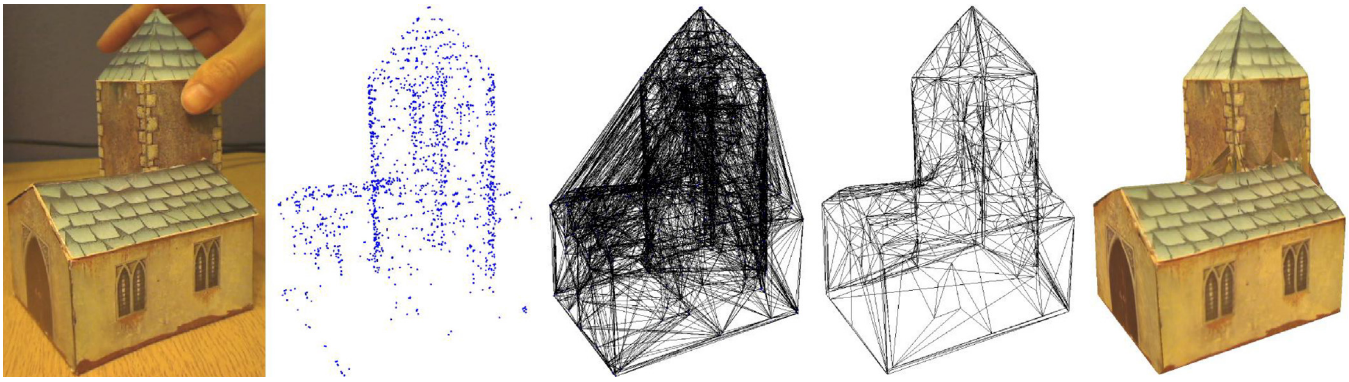


Fig. 2. The ProForma scanning system, reprinted with permission from (Pan et al. [7]). A video showing the scan development can be seen at: http://www.youtube.com/watch?v=vEOmzjlmsVc.

camera and stationary object. Fig. 2 shows the sequence of activities used to generate a 3D model.

The model build process for the system takes around 1 min to complete and feedback is given to the user throughout the process so that he can see what has been done so far and what areas need further scanning. Rapid construction of the model is enabled by using a novel probabilistic tetrahedron carving algorithm which uses the visibility of observed features to quickly create a surface model of the object. Parts of the object that are occluded from the camera are shown in red so that the user can take corrective action to complete the views. In a search environment it is not always necessary to complete the scan however as the user only needs to build a model that is 'good enough' for use as a query. No assumptions are made about the object or its shape, however, some texturing is necessary to provide known points. For metallic objects such as those of interest in this paper this meant marking the object with pens but temporary stickers or labels can also be used.

Use of the system starts with the user showing the object to the camera and the software begins tracking the object. A two threaded keyframe based system is used, as described by [8], whereby separate threads are used for tracking and for reconstruction. Smooth continuous motion of the object is best, with the video sequence providing small distances between points both temporally and spatially which is used for tracking but which provides little 3D information. The tracking thread consists of three trackers, a robust point tracker that follows transient features with 3D location from frame to frame and which is robust to large motions, a second tracker which suffers from less drift, and a 2D tracking function.

The reconstruction thread creates a rendered 3D model from a list of landmarks, keyframes and keyframe camera poses. Because

Table 1
Typical metrics used in generation of shape signature.

Informal descriptor	Definition	Key
Compactness	SA/PA	CHA – convex hull surface area
ConvexHullCompactness	SA/CHA	CHCoG – convex hull centre of gravity or centre of area (assuming constant density)
Crinkliness	PA/PV	CHV – convex hull surface volume
Packing	SV/PV	Dmax – maximum identifiable dimension
AspectRatio0	Lp1/Lp3	Dmin – minimum identifiable dimension
AspectRatio1	Lp1/Lp2	FC – number of facets in triangulated mesh
AspectRatio2	Lp2/Lp3	L1 – shortest distance to a point wrt first principal axis
XYAspectRatio	Lx/Ly	L2 – shortest distance to a point wrt second principal axis
XZAspectRatio	Lx/Lz	L3 – shortest distance to a point wrt third principal axis
YZAspectRatio	Ly/Lz	Ld
SurfaceArea	PA	Lp1 – part length along first principal axis
ConvexHullSurfaceArea	CHA	Lp2 – part length along second principal axis
Volume	PV	Lp3 – part length along third principal axis
ConvexHullVolume	CHV	Lx – part length along *X* axis
DiagonalLength	Ld	Ly – part length along *Y* axis
SmallestDim	Dmin	Lz – part length along *Z* axis
MiddleDim	(Dmax – Dmin/2) + Dmin	PA = part surface area
LargestDim	Dmax	PCoG – part centre of gravity or centre of area (assuming constant density)
XDim	Lx	PV = part volume
YDim	Ly	SA = surface area of sphere with SV = PV (or CHV)
ZDim	Lz	SV = volume of sphere that minimally bounds part
CentreOfAreaRadius	PCoG	
Principal Moment of Inertia0	Sum(L1^2)	
CentreOfAreaRadiusConvex Hull	CHCoG	
Principal Moment of Inertia2	Sum(L2^2)	
Principal Moment of Inertia1	Sum(L3^2)	
Spikeness0	Sum(L1^4)/(L1^2)	
Spikeness1	Sum(L2^4)/(L3^2)	
Spikeness2	Sum(L3^4)/(L3^2)	
FaceCount	FC	

it builds full 3D meshes this method does not require solving the problem of generating 2D views in order to match them with 2D images as described in Refs. [9–11]. As is shown in Fig. 2 a point cloud is created which is then converted into a mesh through a Delaunay tetrahedralisation. Tetrahedra are then carved away based on visibility and probabilistic carving algorithm is used to smooth the resulting surfaces of the model. Finally, textures are added to the model. For the purposes of search however the carved tetrahedral mesh is sufficient to calculate parameters that characterise the object.

On completion of the scan phase the built model is read into the ShapeSpace package for analysis. This involves calculating a number of characteristics of the sample part and using these as a query to search through the part network(s) that are thought to contain the target part. There are multiple methods that have been developed by the research community to characterise shapes [2,3] and these have been used to judge the similarity between parts in databases. Some methods rely on recognising features and on the distribution of these. Whilst these methods appear to be good for fast general purpose searches they do require that features can be identified and they are criticised for being insensitive to feature location within models.

A common alternative to feature methods makes use of spatial maps or functions that typically describe spherical harmonic or a wide variety of methods that are mathematically similar. In general these apply spheres of decreasing size around a voxelised representation of a part and measure the proportion that is on or inside the surface for a given radius. They thus produce signatures that can be compared but they do not, in general, work well with mechanical features such as small threaded holes e.g. skeleton models and other 3D graph representations (e.g. topological graphs, which are similar to feature graphs) are also used as reduced-data models and these can then be compared, but in general use these methods are less sensitive than those previously discussed, especially for typical engineering parts.

The above, and other methods, are typically used for attempts to find closest matching parts. These strategies have had some success, however, they are ultimately limited because they are usually evaluated against some concept of how good they are at recognizing similarity. Since there can be no standard definition of what similarity actually is then there can be no technique that is superior to others except in a practical sense of how well it meets the users' expectations in a particular application. In different contexts users will often have a different concept of what is meant by similar, i.e. similar in what way?

In the approach used by the authors a flexible method is adopted which uses multiple methods of shape characterisation and aggregates these in a way that can support the concept of different types of similarities. The system can therefore be tuned to be more sensitive to some measures and hence be better in particular application areas, e.g. sheet metal or extruded shapes than any general technique might. The system can also adopt methods for partial representations of parts as described in Ref. [12].

3. ShapeSpace

The ShapeSpace program works by initially crawling through a database of parts which might be represented in almost any CAD format and produces STL meshes of these. A wide range of parameters are calculated (in general use 30 different values) and these form the signature of each part. Some of the measures are relatively simple such as the aspect ratio but others are more complex or make use of specific commercially protected algorithms.

Using a 30 entry shape characterising signature allows the use of many previously developed algorithms that are described in the literature. The exact choice used is selected for an individual part environment. This approach also allows the system to readily adopt new measures that can be developed for specific applications. A typical list of measures in use is shown in Table 1.

This table gives a brief description of the individual measures that are used to form the shape signature for any individual 3D model. Thus the shape signature S is a vector of these quantities as follows:

$$S(i) = \{w_1M_1 + w_2M_2 \; + \ldots\ldots\ldots\ldots\ldots\ldots . w_nM_n\}$$

where M_j is a shape descriptor as shown in Table 1, W_j is a weighting factor between 0 and 9, i is the part identifier and n is the total number of shape descriptors in use.

There is clearly overlap between some of these measures, however because they can be pre-calculated, there is little cost in generating them. A principal component analysis (PCA) can be executed for a particular part collection to estimate the extent of this overlap. For example Fig. 2 shows the results of a PCA carried on the use of the measures in a collection 250 CAD files that were generated for various mechanical machine designs (Fig. 3).

The system generates a shape signature for each part stored in the database and when a query is entered it generates a new signature for that part ($S(q)$). A pseudocode version of the search strategy can be as follows:

```
setup{
for each model $i$;
    generate shape signature $S(i)$;
next $i$;
cluster parts according to $k$-means}
loop{
given a query part $q$;
    generate shape signature $S(q)$;
    do until $q$ = target part;
      select cluster for $q$;
      do until screen is full;
        select nearest neighbours;
      allow user to select best guess at target part}
```

Tests with users have shown that in the 'virtual warehouse' 3D environment that the parts are presented in, 256–512 models can readily be viewed and understood. The parts are displayed most likely first (in banks of 25 – see Fig. 5) and then less so the further back they are on screen.

There are various strategies that can be used to generate clusters and these can even be mixed to form complex networks of

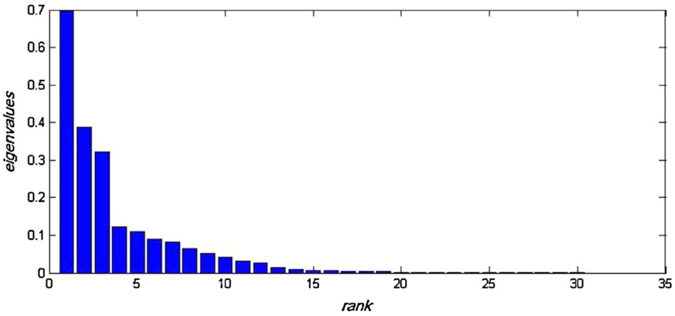
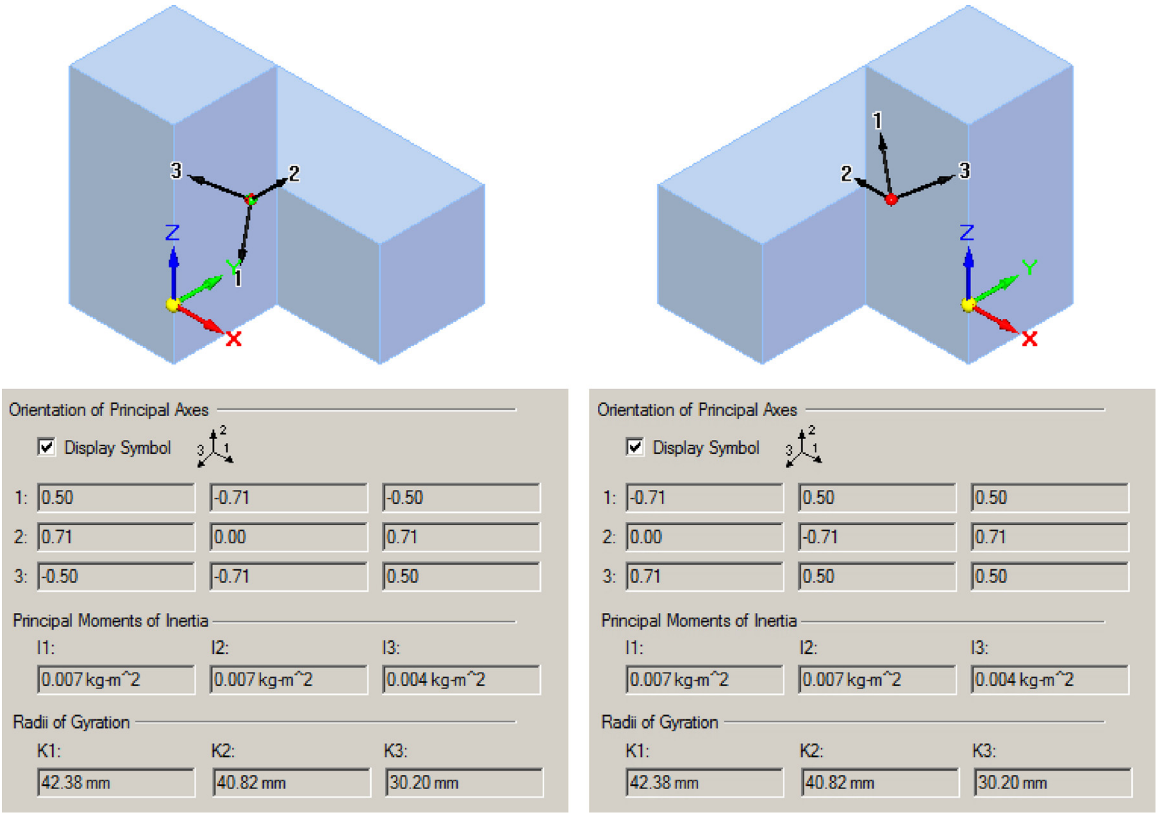


Fig. 3. Results of PCA showing the influence of the various eigenvalues.



Orientation of Principal Axes

☑ Display Symbol

1:	0.50	-0.71	-0.50
2:	0.71	0.00	0.71
3:	-0.50	-0.71	0.50

Principal Moments of Inertia

I1:	I2:	I3:
0.007 kg-m^2	0.007 kg-m^2	0.004 kg-m^2

Radii of Gyration

K1:	K2:	K3:
42.38 mm	40.82 mm	30.20 mm

Orientation of Principal Axes

☑ Display Symbol

1:	-0.71	0.50	0.50
2:	0.00	-0.71	0.71
3:	0.71	0.50	0.50

Principal Moments of Inertia

I1:	I2:	I3:
0.007 kg-m^2	0.007 kg-m^2	0.004 kg-m^2

Radii of Gyration

K1:	K2:	K3:
42.38 mm	40.82 mm	30.20 mm

Fig. 4. Principal axes calculation for mirror image parts showing how orientation results differ between two parts that can be thought of as right and left handed.

parts but standard measures based on Manhatten distance are found to be useful. This method will work well where parts are geometries vary ‘evenly’ across the search space. A problem that is frequently encountered is that of generating false positive chiral parts, finding the left hand version of a right handed part or vice versa. Chiral parts have been found to be very common in automotive, aerospace and many other industry sectors and differentiating between mirror images of parts can be exceedingly complex (we have not found a general solution). In practice the number of false positives can be reduced substantially in nearly all

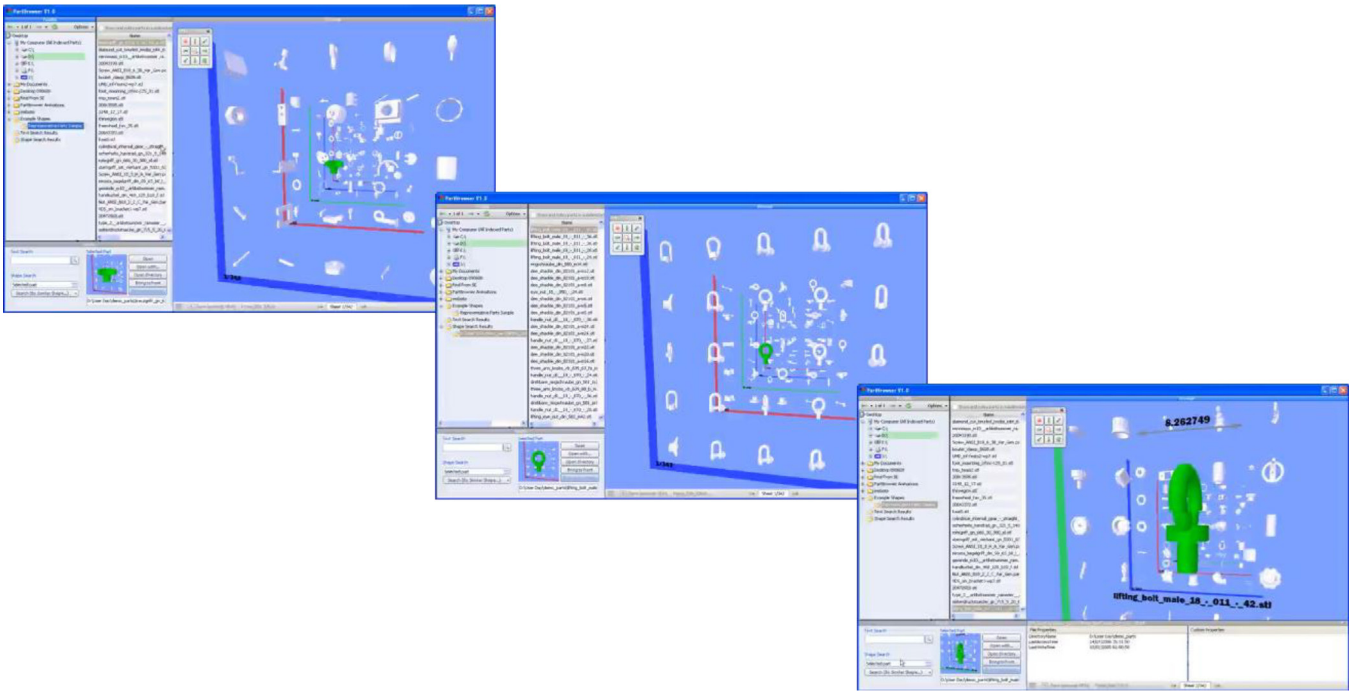


Fig. 5. A typical search sequence using a ShapeSpace. An animation of this can be seen at: http://www.youtube.com/watch?v=YeW7vnaPk7k.

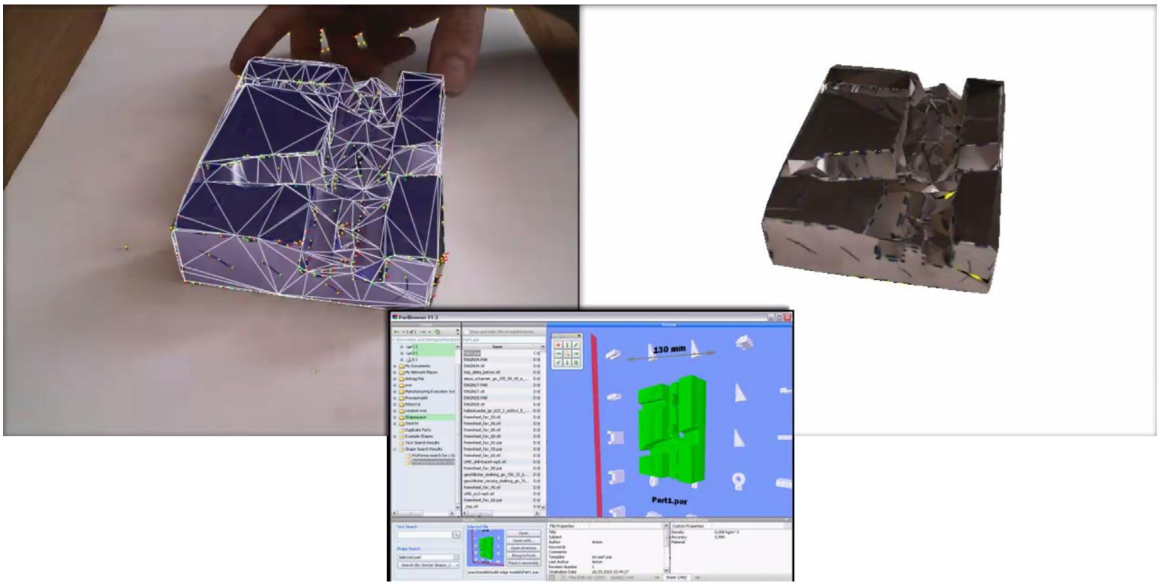


Fig. 6. Integrated scan and search.

cases by making use of the generation of principal axes. Fig. 4 shows the calculation of principal axes for two simple mirror image parts.

If the direction of the first two principal axes for each part are mapped onto each other it is usual to find that the direction of the third axes are opposing and thus chiral parts can be detected. The choice of whether these parts are displayed or not can be left to the user. This problem is particularly severe when duplicate searching is being undertaken rather than simple single part search.

The aim of the search strategy described is not simply to find the best match for a part in one step but to produce a large number of suggestions to the user that represent the search space in a way that allows him to navigate based on his idea of similarity, similar because a part is ‘wavy’ or similar because a part is long and thin. This means that the idea of closeness is not simply based on an aggregated measure of all the parameters. Suggestions are presented to the user on the basis of globally similar parts (from aggregated measures) but also on the basis of parts that are only similar in specific or limited domains. In this way user intent can be employed to guide the search process. The suggested parts or links are given to the user in a 3D warehouse format and the user can easily manipulate the screen to move forward or back through suggestions. Fig. 4 shows a series of screen captures that show a typical search process through a series of screens from Shape-Space.

The example given in Fig. 5 shows a common search through 3 screens. The user starts in this case with an initial screenful of suggestions and picks one most similar candidate to the one being searched for. The part is identified on the 2nd screen and its details given, in this case from a database of 40,000 parts. Longer searches do take place and larger databases have been used but with positive manipulation of the search algorithm, as outlined earlier, and with good suggested parts being offered to the user, based on judicious clustering, it has been found that even in databases of 80,000 plus parts the typical search length is usually around 4 at most and almost always less than 6. Exact statistics do not exist because the databases in use are constantly changing and the purpose and types of search vary continuously.

Searches based on scanned data are typically shorter than those starting with a general screen of parts. In some cases the system will find the required part immediately, however, this is not always possible because scanned parts may be merely similar to those being searched for, for a variety of reasons. Firstly, scanned parts are often worn and damaged and are therefore not perfect representations of the original data version of the part. Secondly, sometimes the target part is not actually the part being used as a query because it is a newer version or replacement part and is therefore ultimately preferred.

4. Integrated scanning and search

By joining the two techniques described so far, the development of a fast system of identifying 3D components is made possible thereby providing the user with whatever linked information is available. A typical search through a network of 3D models (in this case a database of around 40,000 parts) is shown in Fig. 6.

Fig. 6 shows 3 views, the upper left picture represents what the camera sees in terms of recognition of the surfaces of a part presented to it and the triangulation being applied to it. The second view in the upper right shows the generated query model and finally the third bottom centre view shows the part being immediately recognised and identified by the shape matching algorithms.

For subsequent searching and part selection use is made of a network based model of the part database. Although use can sometimes be made of networks built from relations based on common design features, these are unsuitable for the application described in this paper and instead the networks are derived solely from shape measures. The shape measuring and characterisation is based on a crawler that works its way through the database performing calculations and posts the results to a central searchable location. This data reduction means that files can be readily searched without access to the original data, thereby ensuring security of the original data, which may be especially important in cloud based implementations. A wide variety of measures (normally around 30) are calculated and presented as a vector for each part. Common values that are evaluated are the volume, surface area or aspect ratio of the part. There is overlap between some of the measures but the dual aims of assessing similarity but also at times trying to differentiate between parts means that all calculated parameters are currently

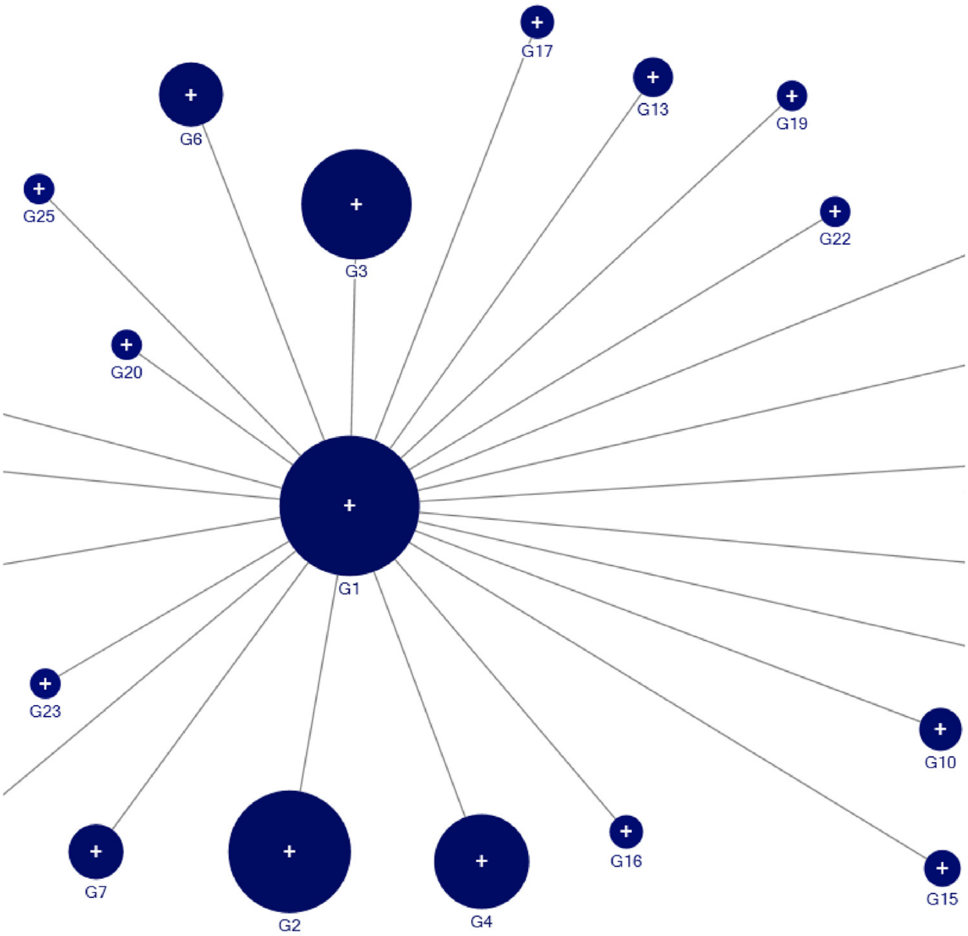


Fig. 7. Network view of part clusters. The groups shown consist of relatively small numbers of parts grouped together by the measures generated from shape signatures.

retained. E.g. in a trial database it was found that the correlation between volume and surface area was close to 0.4 as might be expected from a random population of engineered shapes. It is possible that a principal component analysis (PCA) could reduce the number of measures used however this would save very little time or storage and all of the data can be useful in developing dissimilarity matrices during the clustering phases of the network.

5. Network

Whilst performing post-query searches of a large database the ShapeSpace system does not use a fixed network structure. Instead the links in the network are generated each time the user chooses a part as being similar in some way to the part he is looking for. Thus the parts that are shown on screen may be thought of as the nodes in a network in which every part on screen is joined to every other part on screen. This corresponds to a dynamic clustering approach where all the parts in the database are plotted in a 30 dimensional space. When the user selects a part, the distances from that part to the others can be readily obtained using Euclidean or Manhattan measures and clustering applied so that parts are selected for display to the user on the resulting refreshed view, typically as follows. The closest *n* parts (according to the aggregate total distance) are selected for viewing front and central to the user on screen. Thereafter a further set of *m* suggestion parts are selected if they lie close to the chosen part in one particular dimension (i.e. they may be similar because they have similar relative surface areas despite being dissimilar in other ways e.g.). Finally a few (*l*) parts are chosen at random from parts of the search space not yet sampled.

It is tempting to try to optimise the values of *l*, *m* and *n* that should be chosen to minimise the search, however, this is not possible because the search process is dependent on the navigability of the network rather than any single simple structural aspect of it. Thus only experience with users and the actual network being searched allows some adjustment of the values to be made. However in general the approach does allow the development of a network view that corresponds to complete connectedness, some highly connected nodes and a good degree of navigability for the user. Although these networks cannot be said to be small world they have been found to be relatively efficient from a search perspective. It is also possible in the search tool that previously unused relations can be generated to augment the shape based ones. Previously it was stated that feature data would not be used because it would not be available from the scanned data. Once the user makes a choice to search further there may in fact be such data stored in the database of existing parts and this can then be used. Further information that enhances the searchability of the network could be information regarding features or assembly relationships. Fig. 6 shows a simple network view of a part database with suitable clustering applied based on a dissimilarity matrix Fig. 7.

The size of each node represents the number of parts contained in it. The graph shows how a user can in theory go between any two part models in 2 steps (the geodesic distance) in this database of 500 parts. Larger databases of tens of thousands of parts exhibit geodesic distances of around 4–6.

As has been pointed out navigability or searchability of a network of parts cannot be formally defined because it is dependent on user skill and every search is for a particular purpose and therefore uses different user domain knowledge. In general however it has been found that users are likely to only search for a given amount of time but the system described successfully allows the user to readily find parts with an upper limit of around 5 clicks or 4 choices for well over 90% of cases in large databases of real company data where 80,000 parts have been used.

Although the work and examples reported here was aimed primarily at applications in engineering product management there a considerable uses for this technology in other areas, e.g. medicine, the arts [13] and archaeology [14].

In the setup used in this work, it typically took around 30 s to 1 min to generate sufficient data through laser scanning or fast camera exposure to generate a suitable model for search. It would be useful to further integrate the two systems so that the triangulated data could be continuously sent to the search engine so that the user would become aware as soon the engine found a match in real time, thus minimising the generation of any redundant scanned data.

6. Conclusions

The paper has described a successful method of locating engineering parts in real world databases of 80,000+ parts and which combines various query building methods with a shape browsing strategy. We use as a measure of success the fact that 90% of parts can be located within 6 'clicks' and that computation time is not an issue. The system successfully combines existing and novel methods of shape description in a composite weighted vector that allows flexibility and can be readily adapted to new environments. We argue that search strategies must be able to be customised to specific collection types. Also presented is a practical means of recognising and using (or removing) data relating to left or right handed parts.

7. Equipment

In order to generate the models, searches and graphs presented in this paper a number of tools were adopted as follows. ProFORMA was written as a bespoke application in Linux with C++. Similarly the ShapeSpace system was written in C++ and C# under Windows. The test models were taken from several industrial sites. Other programs were written to clean up and format the data so that network analysis tools could be used to further condition the results and draw the networks. Pajek was used for most of the analysis tasks with NodeXL also being employed, particularly for drawing the network.

References

[1] H. Rea, R. Sung, J. Corney, D. Clark, N. Taylor, Interpreting three-dimensional shape distributions, Proceedings of the Institution of Mechanical Engineers Part C: Journal of Mechanical Engineering Science 219 (June) (2005) 553–566.

[2] J. Tangelder, R. Veltkamp, A survey of content based 3D shape retrieval methods, Multimedia Tools and Applications 39 (2008) 441–471.

[3] A. Cardone, S.K. Gupta, M. Karnik, A survey of shape similarity assessment algorithms for product design and manufacturing applications, Journal of Computing and Information Science in Engineering 3 (2003) 109–118.

[4] C.-F. You, Y.-L. Tsai, 3D solid model retrieval for engineering reuse based on local feature correspondence, International Journal of Advanced Manufacturing Technology 46 (2010) 649–661.

[5] J. Bai, S. Gao, W. Tang, Y. Liu, S. Guo, Design reuse oriented partial retrieval of CAD models, Computer-Aided Design 42 (2010) 1069–1084.

[6] D.E.R. Clark, J.R. Corney, F. Mill, H.J. Rea, A. Sherlock, N.K. Taylor, Benchmarking shape signatures against human perceptions of geometric similarity, Computer-Aided Design 38 (September) (2006) 1038–1051.

[7] Q. Pan, G. Reitmayr, T. Drummond, ProFORMA: probabilistic feature-based on-line rapid model acquisition, Presented at the 20th British Machine Vision Conference, London, UK, 2009.

[8] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces, Presented at the Int'l Symposium on Mixed and Augmented Reality, 2007.

[9] T.F. Ansary, J.P. Vandeborre, M. Daoudi, A framework for 3D CAD models retrieval from 2D images, Annales des telecommunications (Annals of Telecommunications), 60 (2005) 1337–1359.

[10] Y. Wang, R. Liu, T. Baba, Y. Uehara, D. Masumoto, S. Nagata, An images-based 3D model retrieval approach, Lecture Notes in Computer Science 4903 (2008) 90–100.

[11] L. Liang, W. Hanzi, C. Tat-Jun, D. Suter, Z. Shusheng, Retrieving 3D CAD models using 2D images with optimized weights, in: 3rd International Congress on Image & Signal Processing (CISP), 01/04/2010, 2010, p. 1586.

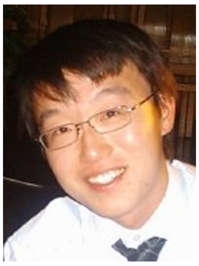
[12] I. Cheuk Yiu, S.K. Gupta, Retrieving matching CAD models by using partial 3D point clouds, Computer-Aided Design & Applications 4 (2007) 629–638.

[13] S. Philipp-Foliguet, M. Jordan, L. Najman, J. Cousty, Artwork 3D model database indexing and classification, Pattern Recognition 44 (2011) 588–597.

[14] A. Itskovich, A. Tal, Surface partial matching & application to archaeology, Computers & Graphics 35 (2011) 334–341.

Frank Mill is chartered engineer and works as a senior lecturer at The University of Edinburgh. He graduated with a B.Sc. (honours: first class) Technology with Industrial Studies in 1983 from Napier College and with a Ph.D. in Shape Optimisation in 2004 from the University of Edinburgh. He has worked with Redcastle Systems and ShapesSpace Ltd. on technical aspects of CAD. He is a co-founder and director of ShapeSpace Ltd and has carried out extensive research, consultancy and teaching in CAD.

Andrew Sherlock is the CEO of ShapeSpace Ltd and Technical Director of Actify (UK) Ltd. He graduated with a first class B.Eng. honours degree in electrical and mechanical engineering (1995) and with a Ph.D. in shape optimisation (2004) from the University of Edinburgh. He has held positions in various companies involved with optimisation and CAD/PLM work and is a co-founder of ShapeSpace Ltd.



Qi Pan is a senior engineer at Qualcomm Austria Research Centre GmbH. He graduated in 2007 with a B.A./M.Eng. with distinction in electrical and information sciences and with a Ph.D. in computer vision in 2010 from the University of Cambridge.

Esme Anderson is currently pursuing a Ph.D. in computer aided design at the University of Edinburgh. She graduated with an upper second class Master of Engineering in mechanical engineering in 2011 from the University of Edinburgh.

Detection of design reuse in 3D CAD using network analysis

Esmé Anderson*, Dr. Frank Mill**

Plagiarism isn't limited to the written word, some research has focused on non-text plagiarism (Blythman et al. 2007), but little has been done (Houjou 2013) to detect and prevent copying in 3D Computer Aided Design (CAD) courses.

With more universities and schools teaching 3D design it is of paramount importance that students learn to present authentic work and not fall into the trap of design plagiarism (Martin et al. 2006).

This form of academic dishonesty is particularly difficult to detect in assignments and project work from CAD courses, where students use specialist software to design and create models of many different objects, from shoes and circuit boards to houses and cars. With 3D CAD modelling for engineers it is nearly impossible to spot use of previous work in hand-ins. Large university classes make it unrealistic for a marker to identify plagiarised material.

To this end an investigation into plagiarism detection in mechanical engineering CAD assessments has been undertaken. A new and unique method utilising shape similarity technology and network analysis has been employed to detect similar designs within a class.

Real data was taken and anonymised from a class at the University of Edinburgh. 'Cheating students' were created and added to the collection. These 'cheats' were designed to simulate direct copying or theft of another's work from one or more students.

The CAD files were then analysed for shape similarity and the results fed into a graph database via NodeXL (free open-source software). This generated a network graph where the vertices represent a student's submissions and the edges show similarity between them.

Utilising the metrics provided by graph theory, methods to identify the 'cheats' in the large collection of files were explored.

Mapping these graph measurements to visual properties, we created an illustrative network, which highlighted the 'cheats'. (See Fig. 1)

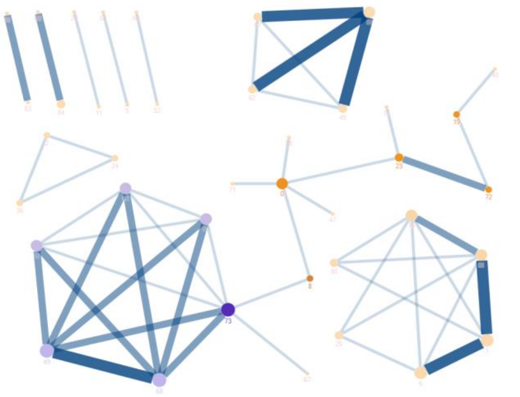


Figure 1: Network Graph of Students’

Here you can see the large opaque nodes and the many, heavy edges which highlight concerning situations.

It was found that the 'cheating students' could be located by some key metrics: A large betweenness centrality, which indicates a highly linked student and a high degree with heavy edges, indicating multiple strong links with other students designs.

Initial result show it would be possible to highlight students of concern to a marker for further investigation, making the task of verifying original work much easier and less time consuming.

Experiments are currently underway with other sets of CAD submissions to verify these results.

This paper concludes that network analysis techniques have the potential to effectively and reliably highlight possible plagiarism in 3D CAD files.

Blythman, M., Orr, S. & Mullin, J., 2007. Reaching a consensus: plagiarism in non-text based media. *London College of Communication, University of the Arts London, London.*

HOUJOU, K., 2013. Development of Evaluation Methodology for Appropriate 3D-CAD Practice on Mechanical Design Education. *Journal of JSEE*, 61(2), pp.2_7–2_11.

Martin, I., Stubbs, M. & Troop, H., 2006. Weapons of mouse destruction: a 3D strategy for combating cut-and-paste plagiarism using the JISC Plagiarism Advisory Service.

*e.anderson@ed.ac.uk, The University of Edinburgh

**The University of Edinburgh

THE UNIVERSITY *of* EDINBURGH

New things come from our weak ties!

Better Connected: 3D CAD Networks

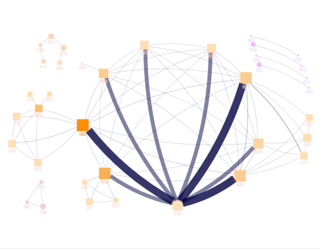
Networks are prolific in our lives,
but rarely used in Mechanical Engineering applications.

This research investigates mapping and analysing **CAD** directories
using the **shape similarity** of 3D design parts

Network graphs are made up of vertices (points) and edges (links).
Follow the directional edges to discover more about this research.

CAD Plagiarism

Many schools and universities teach **CAD**, but there is currently no easy and reliable way to authenticate students work. **Turnitin** makes it possible to detect plagiarism in written work, but when students are submitting 3D design files for assessment, markers cannot identify if they are copied or original.


Network showing anonymised student submissions

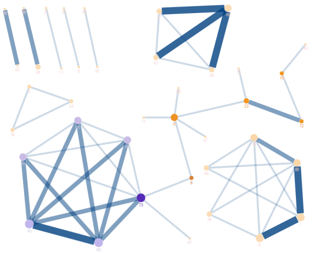
Investigation into **detecting authentic work** began utilising networks formed from students CAD submissions. In these networks the vertices represent students, and the edges show those how their work is linked by shape similarity.

We created '**cheats**' who had directly copied or edited other students files and analysed the resulting network.

In theory this should show us how closely linked students' work it. If one has directly copied another there should be a clear link, just as if one has taken files from several students.

Using NodeXL (free, open-source software) the data was analysed and measurements taken. This software allows the metrics to be translated into visual properties and graphs to be produced as shown (above and right)

Translating the **metrics into graphics** allows us to visually assess the results and locate the 'cheats'. In these graphs heavy edges indicate students with many similar files while large, opaque vertices indicate students whose files are similar to many others in the groups.


Network highlighting created plagiarising student '0'

Initial results show that students with many similar files can be easily located in the database using graph metrics. There are two key metrics that highlight problem students to us: **betweenness centrality and edge weight**.

It is not possible to know if these are true plagiarism instances without **human interpretation**. To this end it may be possible to use these results in detection software, so a marker would be able to assess a small number of highlighted students of concern.

With more results it should be proved that with a little human interaction 3D CAD plagiarism can **easily and reliably** be detected using this method.

Business Card holder

Network with me

Esmé Anderson
e.anderson@ed.ac.uk

Supervisor: Dr. Frank Mill
Industrial Sponsor: ShapeSpace Ltd.

ShapeSpace

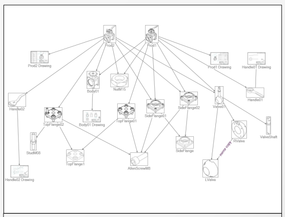
EPSRC
Engineering and Physical Sciences Research Council

Motivation

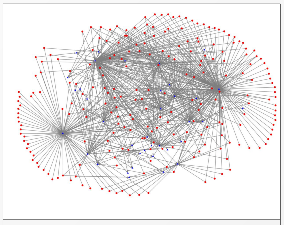
Shape similarity technology is already very advanced, 3D CAD models can be assessed by shape to determine how similar they are to one another. A database of this information allows design engineers to search for and use previous designs. Networks can create graphs of shape similarity, but there has been no assessment of their usefulness.

This work continues the research of Dr. Frank Mill.
Using networks in mechanical engineering we are asking:

- **Can CAD databases be more "searchable"?**
- **Can 3D plagiarism be detected?**
- **What do graph metrics tell us about CAD databases?**


'Assembly' structure shown as network graph FM

Below is a network, created from a collection of 250 engineering CAD parts. This network is bimodal; the red vertices are parts and the blue vertices are features. The edges between them show the relationship, illustrating which parts have which features.


Network graph of Edinburgh benchmark parts. FM

Networks can model any relationship between vertices. Another logical type of CAD relationship to model is an assembly tree structure (shown above). Here the edges are directional, showing us how a final design is composed of single parts.

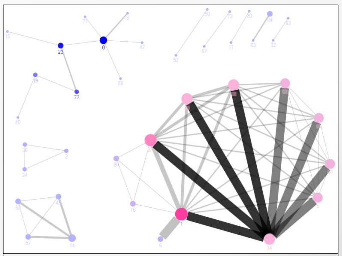
Using these networks we can begin to measure and assess the graphs we have created.

The measurements we take provide us with data about the size, shape and connectedness of the network. Utilising these data we are assessing collections of CAD parts.

It is important for us to redefine what these data mean in this application. E.g.:
Average Geodesic Distance is the average minimum number of edges between any 2 vertices. If the network is CAD models linked by shape, average geodesic distance represents how many steps it takes to search for a model from another, on average.

Results so far...

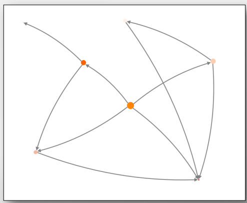
From initial results it appears networks and graph theory can be used to highlight plagiarism instances within 3D CAD model design submissions. This must be verified with further data analysis. Below is an example of a collection of students work, which highlights students of concern.


Network highlighting created cheating student '0'

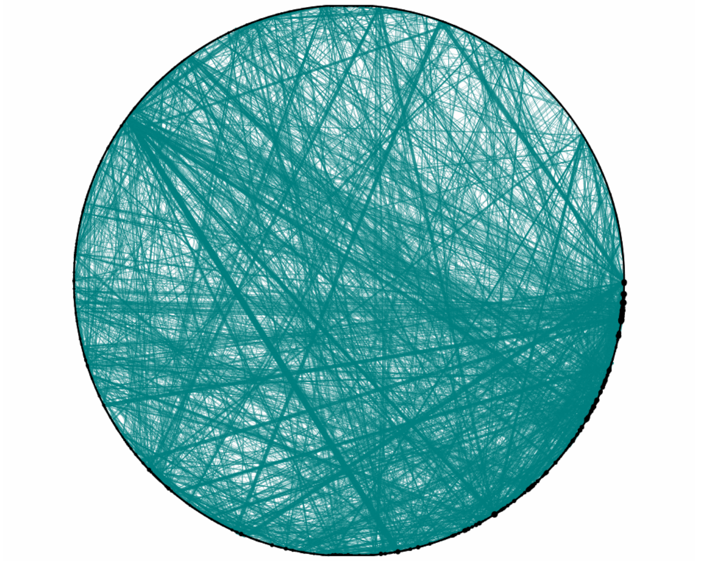
How could better connections affect your work?

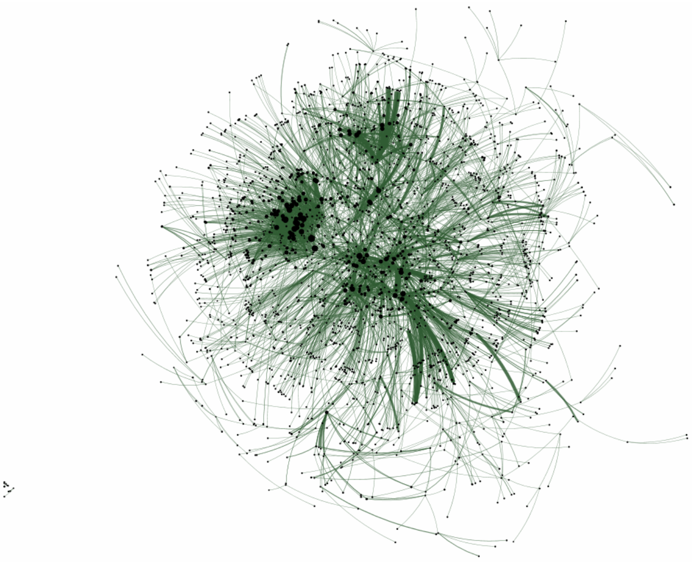
This Network: Metrics of this poster

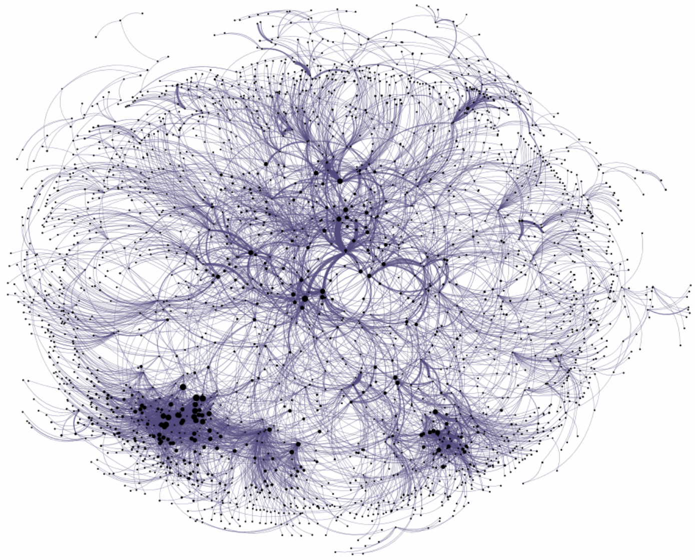
Figure shows network diagram of this poster

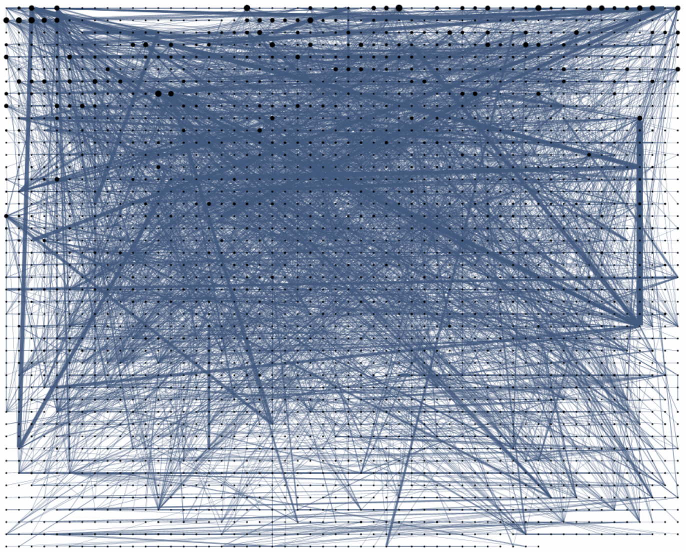


Graph Density: 0.248
Average Geodesic Distance: 1.51
Max Geodesic Distance: 4
Average Betweenness Centrality :4.571
Average Closeness Centrality: 0.099
Average Eigenvector Centrality: 0.143
Average Clustering Coefficient: 0.262









Seeing things differently

Unlocking engineering design secrets

Engineering design companies are at the cutting edge of technology, but rarely have a clear picture of their data. Sleek and impressive 3D models don't show the structure of the data behind them and the complex data management systems utilised do not empower companies with realistic knowledge of how profits relate to components.

Applying network theory to the data already held in these advanced systems has been shown to illuminate previously unrevealed information that impacts engineering design.
These network graphs show four different views of the same 15,000 Computer Aided Design (CAD) models. Some are full assemblies, others are sub-assemblies and then there are components. The relationship modelled is 'contains' illustrating which assemblies contain which components.

These novel visualisations of design data show formally impossible and unseen insights, which can guide companies to reduce the unnecessary variety in their products. Utilising the metrics and analysis tools provided by network theory unveils even more, showing which are the key components in designs, which components are less used and even highlights those components which are so similar they should be amalgamated.

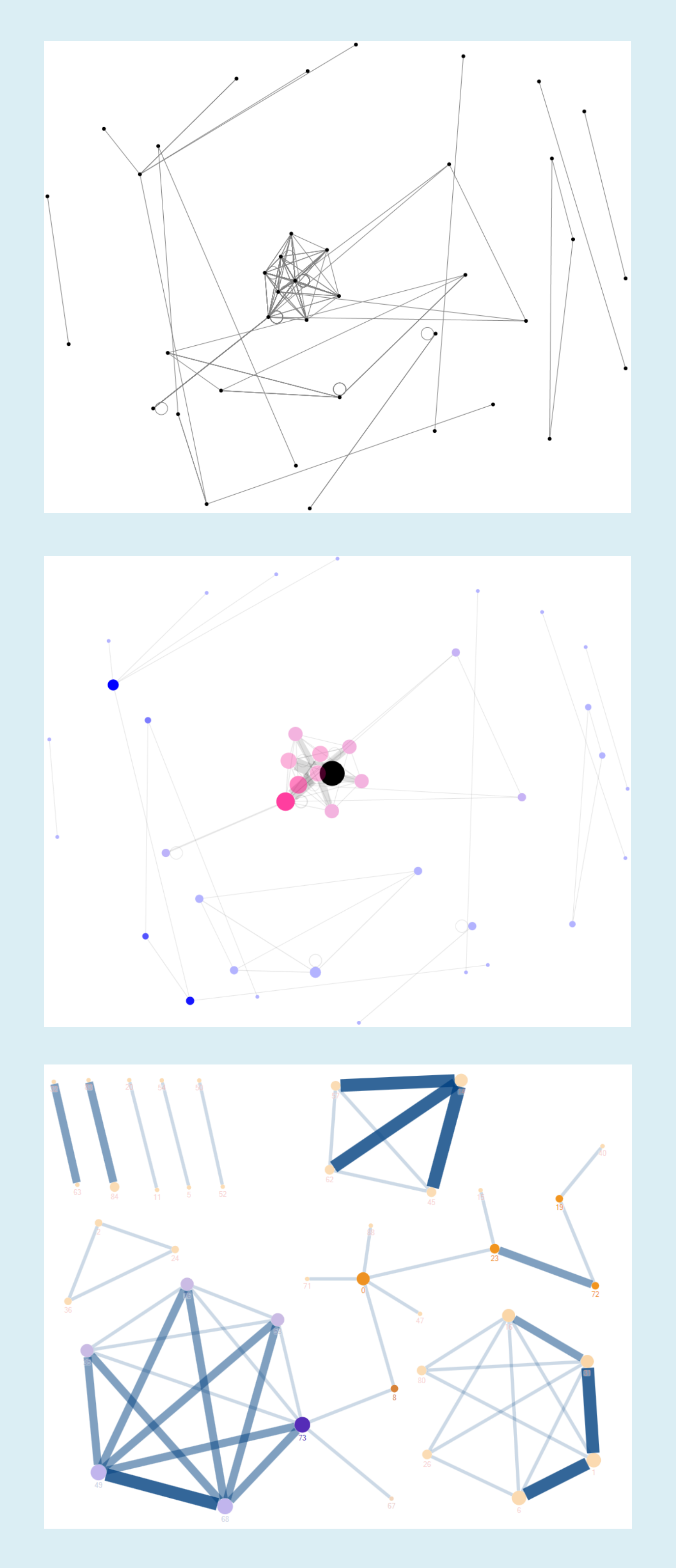
Coupling this new information with cost, manufacturing, orders, and customer data gives companies capability to decide whether their designs and product are valuable and which should continue to be made, based on profitability and sales.

While these network graphs provide all this information, they are also visually stunning, providing unique and hitherto unseen visualisation of company data akin to the advanced 3D models these companies pride themselves on.

Esmé Anderson: e.anderson@ed.ac.uk - Dr. Frank Mill: Frank.Mill@ed.ac.uk – ShapeSpace Ltd: shapespace.com

ShapeSpace

EPSRC
Engineering and
Physical Sciences
Research Council



Seeing things differently

Visualising Plagiarism: Detecting 3D CAD cheats

Plagiarism is increasingly problematic in education. It is especially difficult to trace in non-text submissions. For text based assignments there is TurnItIn, a plagiarism detection software, but there is no such detection aid for image, code or 3-Dimentional based assignments.

In engineering Computer Aided Design (CAD) is a key part of education, where students are taught to design, model and render 3D products. CAD assessments are particularly difficult to detect plagiarism in as each student submits a large number of files with many attributes, some of which may similar due to the assignment specifications.

One theory we have investigated to address this is the use of network theory, which is illustrated above.

To create this visualisation data was collected from an engineering design class (Computer Aided Engineering 3) at the University of Edinburgh. A cheating student was fabricated, all files were compared by shape and a network was generated from the results. The nodes in the network represent a single student's collection of work and the edges show a 'similar by shape' relationship.

The series of images above show the progression of data visualisation. The first is the raw data in a Fruchterman-Reingold layout. The second image has calculated metrics applied as visual properties and the third is a hand-arranged layout, allowing for clear interpretation of the results.

In the last image the fictional cheating student can clearly be identified. This student '0' had the highest "betweeness centrality" and the smallest "clustering coefficient". When this cheat was created, it was decided they should have 'stolen' files from several different students, edited some and not changes others. This realistic model of plagiarism was detectable using network analysis.

Further work is underway to prove this remains true for other data sets and we aim to show that analysing CAD data in this way can highlight plagiarism; something which has never been possible before.

Esmé Anderson: e.anderson@ed.ac.uk - Dr. Frank Mill: Frank.Mill@ed.ac.uk – ShapeSpace Ltd: shapespace.com

ShapeSpace

EPSRC
Engineering and Physical Sciences Research Council

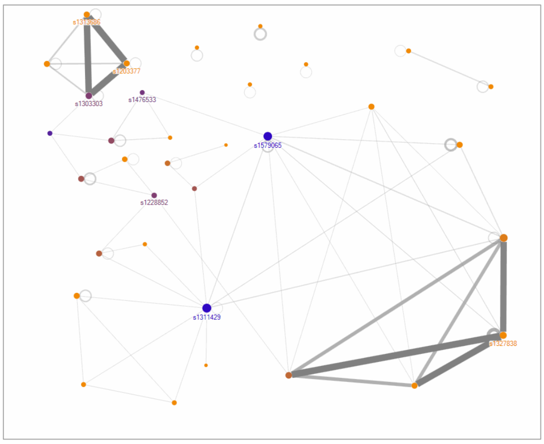
Appendix B

Originality report

Esmé Anderson, University of Edinburgh

Originality report CAD3F 2015-2016

Following the 3rd year hand in for the CAD course, 58 students work was analysed. The ShapeSpace legacy indexer processed 45 students work; 1035 of the 1710 CAD files handed in were passed to the duplicate analyser. Using NodeXL and the developed technique the following graph has been produced.



There were 36 students who were found to have similar work. The results above suggest several students are likely to have unoriginal work in their file and these are s11228852, s1303303, s1311429, s1476533 and s1579065, with s1311429 and s1579065 being indicated as the most likely to have handed in unoriginal work. Also s1303303 is of particular concern due to their position in the network. The students are represented by the largest, bluest nodes in the above graph. Also of note are students s1203377 and s1313686, who work has strong similarities to s1303303. Therefore if student s1303303 is found to have unoriginal work, it may be the case that s1203377 and s1313686 do also. Also of note is s01327838, however the large self-loop here may be the cause of the strong links it seems they have with other students.

Conclusion

This report suggests further investigation of the work belonging to students s11228852, s1303303, s1311429, s1476533 and s1579065.
Students s1203377 and s1313686 work should be further scrutinized if s1303303 is found to have unoriginal work.

Exam diet 2015-2016

Appendix C

Data Visualisation

In this work network visualisation is discussed, and data visualisation is touched on. In this appendix background information for data visualisation will be presented, alongside other images produced during this research. These network diagrams were created alongside various performed analysis and are considered interesting, but do not fit into a relevant section.

Background of Data Visualisation and Infographics

Data visualisation is not limited to network visualisation. Visualisation of data is now so common it is part of daily life without being immediately obvious. One key example is advertising, which is more notable as technology has provided further opportunities for information communication. There are many questions surrounding the efficacy of advertising and while research has not conclusively proved how much of this information people absorb, there have been noted developed phenomena, including peoples ability to 'tune out' advertisements [73]. While this is an issue for marketers, it does not affect the creation or use of data visualisations to communicate with the public.

As well as a popular choice for advertisers, it has become common practice for researchers within science and engineering to use data visualisation to present information, making use of graphs, charts and infographics. Within universities courses are run on how academics can more effectively communicate their research to colleagues, peers and the public. One such course run at the University of Edinburgh by Iain Davidson focuses on the importance of visuals to communicate clearly [53]

Infographics combine words and images to visually communicate. They incorporate diagrams, charts, graphs, signs, maps or pictures with text to strengthen the viewers understanding of the content. In her masters thesis Mol (2011) highlighted the difficulty in finding an official definition of infographics, given the relatively new field has little associated literature or research. While infographics have increased in popularity recently, often suggested as a result of the Internet and social media such as Twitter [30] and despite the scarcity of academic work in the field, infographics are hardly new or novel.

This history and advance of infographics is linked with the development of graph theory; as the methods for visually representing data with graphs were developed, so were the different varieties of charts, and these charts act an important component of many infographics and other data visualisation. It is suggested (by both academics and graphic designers) that the earliest form of informatics were cave paintings, Egyptian hieroglyphics [113, 173], maps (dated as early as 7000BC) [129] and Leonardo Da Vinci's 'Vitruvian Man' (1490) [134].

In 1626 Christoph Scheiner studied the sun and published diagrams depicting the rotation of the sun [162] after which William Playfair, an engineer and political economist published the first data charts, including the first pie chart in 1786. Widely acknowledged as the first major work to contain statistical graphs, Playfair is also viewed as the inventor of most common graphical forms. He commented on using visualisations to communicate "As the eye is the best judge of proportion, being able to estimate it with more quickness and accuracy than any other of our organs, it follows, that wherever relative quantities are in question [the Line Chart] is peculiarly applicable; it gives a simple, accurate, and permanent idea, by giving form and shape to a number of separate ideas, which are otherwise abstract and unconnected" [146].

Florence Nightingale is the next to have produced a noteworthy infographic. Better known for her role as a nurse, she produced a document depicting the causes of mortality of the British Army during the Crimean War, combining charts and a diagram that was presented to Queen Victoria in order to convince her, parliament members and civil servants that medical care must be improved. Lankow *et al.* argue this proves the efficacy of infographics, as Queen Victoria was persuaded, and would have been unlikely to understand a traditional statistical report [113]

Data visualisation techniques continued to develop, with a marked growth attributed to the Internet [134] and are now an accepted part of modern daily life. Recent noteworthy work has been undertaken in this area by the data-journalist David McCandless, who has produced two prominent books about the visualisation of data; ‘Information is Beautiful’ (2009) and ‘Knowledge is Beautiful’ (2014). In both books McCandless uses infographics to communicate all different kinds of data. He claims to have first started creating the infographics due to feeling swamped by information, finding, in agreement with Playfair, that visualisations were a better way to see and understand it all [126, 128]. In his 2010 talk he extolled the virtues of using data visualisation;

> “By visualising information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you’re lost in information, an information map is kind of useful”. [127]

McCandless’s work and ethos is one example of the varied modern applications of infographics. In his 2014 book ’Knowledge is beautiful’ McCandless even included an infographic entitled ‘What Makes a Good Visualisation’ shown in figure C.1. This image combines many the common elements of infographics, with the concept of ven diagrams to effectively communicate the important elements that must be balanced to produce a good visualsation, while achieving what it claims to communicate. With many infographics already making use of charts and graphs, so it is understandable that networks diagrams could be used as a key element in data visualisation, to effectively communicate with an audience.

The use of data visualisation to communicate is effective and a recognised field, especially in the public eye. As such a small overview shows the importance of images to communicate effectively. With this in mind, the work undertaken in this thesis will seek to explore the uses of network theory, including network visualisation within mechanical engineering design scenarios.

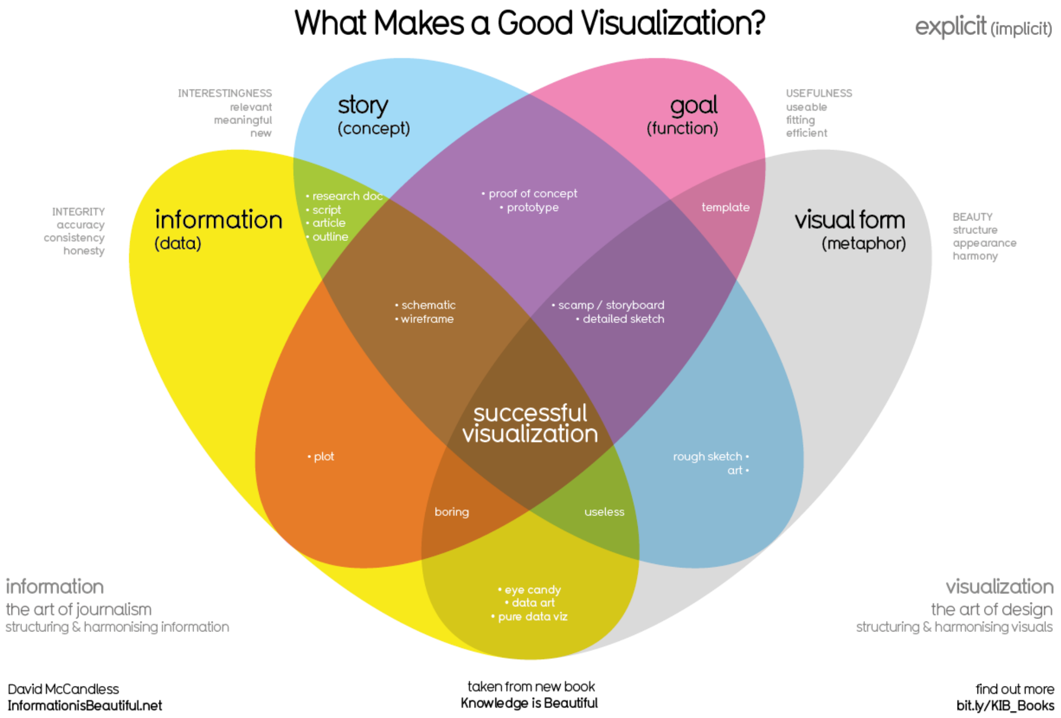


Figure C.1: What Makes a Good Visualisation?

Bibliography

[1] Abbott, L., and van Vreeswijk, C. Asynchronous states in networks of pulse-coupled oscillators. *Physical Review E 48*, 2 (1993), 1483.

[2] Akgül, C. B., Sankur, B., Yemez, Y., and Schmitt, F. Similarity learning for 3d object retrieval using relevance feedback and risk minimization. *International Journal of Computer Vision 89*, 2-3 (2010), 392–407.

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

[4] Aldana-Gonzalez, M. Nexus: Small Worlds and the Groundbreaking Science of Networks; Linked: The New Science of Networks. `http://scitation.aip.org.ezproxy.is.ed.ac.uk/content/aip/magazine/physicstoday/article/56/3/10.1063/1.1570777`, 2003. Accessed: 2015-09-15.

[5] Alemzadeh, K., and Burgess, S. A team-based cad project utilising the latest cad technique and web-based technologies. *International Journal of Mechanical Engineering Education 33*, 4 (2005), 294.

[6] Allen, M. Dematerialised data and human desire: the internet and copy culture. In *Cyberworlds, 2003. Proceedings. 2003 International Conference on* (2003), IEEE, pp. 26–33.

[7] Altfeld, N., Hinckeldeyn, J., Kreutzfeldt, J., and Gust, P. Impacts on supply chain management through component commonality and postponement: A case study. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (2011), vol. 2, Citeseer.

[8] Ariely, D. The honest truth about dishonesty. In *6th International Integrity and Plagiarism Conference* (2014).

[9] Arrish, S., Afif, F. N., Maidorawa, A., and Salim, N. Shape-based plagiarism detection for flowchart figures in texts. *arXiv preprint arXiv:1403.2871* (2014).

[10] Bai, J., Gao, S., Tang, W., Liu, Y., and Guo, S. Design reuse oriented partial retrieval of cad models. *Computer-Aided Design 42*, 12 (2010), 1069–1084.

[11] Bai, J., Luo, H., and Qin, F. Design pattern modeling and extraction for cad models. *Advances in Engineering Software 93* (2016), 30–43.

[12] Barabasi, A. L. *Linked.* Perseus Publishing, 2002.

[13] Barnes, J. A., and Harary, F. Graph theory in network analysis. *Social Networks 5*, 2 (1983), 235–244.

[14] Baron, P., Fisher, R., Tuson, A., Mill, F., and Sherlock, A. A voxel-based representation for evolutionary shape optimization. *AI EDAM 13*, 03 (1999), 145–156.

[15] Bastian, M., Heymann, S., Jacomy, M., et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM 8* (2009), 361–362.

[16] Batagelj, V., and Mrvar, A. Pajek-program for large network analysis. *Connections 21*, 2 (1998), 47–57.

[17] BBCNews. India students caught 'cheating' in exams in Bihar. `http://www.bbc.co.uk/news/world-asia-india-31960557`, 2015. Accessed: 2015-03-20.

[18] BBCNews. Sam Smith: Tom Petty given writing credit for Stay With Me. `http://www.bbc.co.uk/news/entertainment-arts-30997759`, 2015. Accessed: 2015-01-27.

[19] Becker, R. A., Eick, S. G., Miller, E. O., and Wilks, A. R. Dynamic graphics for network visualization. In *Proceedings of the 1st conference on Visualization'90* (1990), IEEE Computer Society Press, pp. 93–96.

[20] Becker, R. A., Eick, S. G., and Wilks, A. R. Visualizing network data. *Visualization and Computer Graphics, IEEE Transactions on 1*, 1 (1995), 16–28.

[21] Berners-Lee, T., Fischetti, M., and Foreword By-Dertouzos, M. L. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* HarperInformation, 2000.

[22] Bertin, J. *Graphics and graphic information processing.* Walter de Gruyter, 1981.

[23] Bespalov, D., Ip, C. Y., Regli, W. C., and Shaffer, J. Benchmarking cad search techniques. In *Proceedings of the 2005 ACM symposium on Solid and physical modeling* (2005), ACM, pp. 275–286.

[24] Biasotti, S., De Floriani, L., Falcidieno, B., Frosini, P., Giorgi, D., Landi, C., Papaleo, L., and Spagnuolo, M. Describing shapes by geometrical-topological properties of real functions. *ACM Computing Surveys (CSUR) 40*, 4 (2008), 12.

[25] Biasotti, S., Giorgi, D., Spagnuolo, M., and Falcidieno, B. Size functions for comparing 3d models. *Pattern Recognition 41*, 9 (2008), 2855–2873.

[26] Biggs, N., Lloyd, E. K., and Wilson, R. J. *Graph Theory, 1736-1936.* Oxford University Press, 1976.

[27] Blythman, M., Orr, S., and Blair, B. Critiquing the crit. *Brighton: Art, Design and Media Subject Centre* (2007).

[28] Blythman, M., Orr, S., and Mullin, J. Reaching a consensus: plagiarism in non-text based media. *London College of Communication, University of the Arts London, London. Available at* (2007).

[29] Bourke, D. Where Search & Discover Solutions Fit in Product Development. `http://www.engineering.com/DesignSoftware/DesignSoftwareArticles/ArticleID/8467/Where-Search-Discover-Solutions-Fit-in-Product-Development.aspx`, 2014. Accessed: 2014-09-18.

[30] Bradshaw, L. The rise of infographics: Lessons from the Social Media Weekend. `http://www.cbsnews.com/8301-504943{\_}162-20064104-10391715.html`, 2011. Accessed: 2016-01-26.

[31] Braiman, Y., Lindner, J. F., and Ditto, W. L. Taming spatiotemporal chaos with disorder. *Nature 378*, 6556 (1995), 465–467.

[32] Bressloff, P., Coombes, S., and De Souza, B. Dynamics of a ring of pulse-coupled oscillators: Group-theoretic approach. *Physical review letters 79*, 15 (1997), 2791.

[33] Broad, S. *Finding Models with a ShapeSpace.* Masters thesis, University of Edinburgh, 2013.

[34] Bromwich, J. Six Degrees of Separation? Facebook Finds a Smaller Number. `http://www.nytimes.com/2016/02/05/technology/six-degrees-of-separation-facebook-finds-a-smaller-number.html?utm{\_}source=pocket{\&}utm{\_}medium=email{\&}utm{\_}campaign=pockethits`, 2016. Accessed: 2016-02-08.

[35] Bronstein, A. M., Bronstein, M. M., Bruckstein, A. M., and Kimmel, R. Partial similarity of objects, or how to compare a centaur to a horse. *International Journal of Computer Vision 84*, 2 (2009), 163–183.

[36] Browning, G. Bob Dylan: Creativity and plagiarism. In *6th International Integrity and Plagiarism Conference* (2014).

[37] Buchanan, M. *Small world: uncovering nature's hidden networks.* Diane Publishing Company, 2002.

[38] Bustos, B., Keim, D., Saupe, D., and Schreck, T. Content-based 3d object retrieval. *Computer Graphics and Applications, IEEE 27*, 4 (2007), 22–27.

[39] Bustos, B., Keim, D. A., Saupe, D., Schreck, T., and Vranić, D. V. Feature-based similarity search in 3d object databases. *ACM Computing Surveys (CSUR) 37*, 4 (2005), 345–387.

[40] Cardone, A., Gupta, S. K., and Karnik, M. A Survey of Shape Similarity Assessment Algorithms for Product Design and Manufacturing Applications. *Journal of Computing and Information Science in Engineering 3*, 2 (2003), 109.

[41] Cartwright, D., and Harary, F. Structural balance: a generalization of heider's theory. *Psychological review 63*, 5 (1956), 277.

[42] Cayley, A. Xxviii. on the theory of the analytical forms called trees. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 13*, 85 (1857), 172–176.

[43] Cayley, A. On the mathematical theory of isomers. *Philosophical Magazine 67*, 5 (1874), 444–446.

[44] Chaouch, M., and Verroust-Blondet, A. Alignment of 3d models. *Graphical Models 71*, 2 (2009), 63–76.

[45] Chu, C.-H., and Hsu, Y.-C. Similarity assessment of 3d mechanical components for design reuse. *Robotics and Computer-Integrated Manufacturing 22*, 4 (2006), 332–341.

[46] Chuda, D., Navrat, P., Kovacova, B., and Humay, P. The issue of (software) plagiarism: A student view. *Education, IEEE Transactions on 55*, 1 (2012), 22–28.

[47] Clark, D. E., Corney, J. R., Mill, F., Rea, H. J., Sherlock, A., and Taylor, N. K. Benchmarking shape signatures against human perceptions of geometric similarity. *Computer-Aided Design 38*, 9 (2006), 1038–1051.

[48] Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., and Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology 62*, 7 (2011), 1382–1402.

[49] Collins, J., Chow, C. C., Imhoff, T. T., et al. Stochastic resonance without tuning. *Nature 376*, 6537 (1995), 236–238.

[50] Crace, J. Here's one I ripped off earlier. `http://www.theguardian.com/education/2007/jan/23/highereducation.uk1`, 2007. Accessed: 2015-01-29.

[51] Culwin, F., and Lancaster, T. Plagiarism issues for higher education. *Vine 31*, 2 (2001), 36–41.

[52] Dankwort, C. W., Weidlich, R., Guenther, B., and Blaurock, J. E. Engineers' cax educationit's not only cad. *Computer-Aided Design 36*, 14 (2004), 1439–1450.

[53] Davidson, I. Writing and Designing Your Academic Posters: School of Engineering, 2014.

[54] De Nooy, W., Mrvar, A., and Batagelj, V. *Exploratory social network analysis with Pajek*, vol. 27. Cambridge University Press, 2011.

[55] Demian, P., and Fruchter, R. An ethnographic study of design knowledge reuse in the architecture, engineering, and construction industry. *Research in Engineering Design 16*, 4 (2006), 184–195.

[56] Dictionary, O. E. Oxford english dictionary online. *Mount Royal College Lib., Calgary 14* (2004).

[57] Economou, I. The problem with plagiarism. *20/20 Design Vision* (2011), 79.

[58] Euler, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae 8* (1741), 128–140.

[59] Falcidieno, B., and Herman, I. Special section on semantic 3d media and content. *Computers & Graphics 35*, 4 (2011), iii–iv.

[60] Fang, R., Godil, A., Li, X., and Wagan, A. A new shape benchmark for 3d object retrieval. In *Advances in Visual Computing*. Springer, 2008, pp. 381–392.

[61] Ferreira, A., Marini, S., Attene, M., Fonseca, M. J., Spagnuolo, M., Jorge, J. A., and Falcidieno, B. Thesaurus-based 3d object retrieval with part-in-whole matching. *International Journal of Computer Vision 89*, 2-3 (2009), 327–347.

[62] Field, D. A. Education and training for cad in the auto industry. *Computer-Aided Design 36*, 14 (2004), 1431–1437.

[63] Filaliansary, T., Vandeborre, J.-P., and Daoudi, M. A framework for 3d cad models retrieval from 2d images. In *Annales des télécommunications* (2005), vol. 60, Springer, pp. 1337–1359.

[64] Findlay, K. An introduction to Network Theory. In *SAMRA 2010 Conference* (Magaliesburg, South Afria, 2010), p. 50.

[65] Flajolet, P., and Martin, G. N. Probabilistic counting algorithms for data base applications. *Journal of computer and system sciences 31*, 2 (1985), 182–209.

[66] Frankland, E. *Lecture notes for chemical students: Embracing mineral and organic chemistry*. John van Voorst, 1866.

[67] Friedkin, N. A test of structural features of granovetter's strength of weak ties theory. *Social Networks 2*, 4 (1980), 411–422.

[68] Gabriel, T. Lines on Plagiarism Blur for Students in the Digital Age. `http://www.nytimes.com/2010/08/02/education/02cheat.html?{\_}r=0`, 2011. Accessed: 2015-01-29.

[69] Gallant, T. B. Where next? Integrity for the "Real World". In *6th International Integrity and Plagiarism Conference* (2014).

[70] Garrett, L., and Robinson, A. Spot the difference! visual plagiarism in the visual arts. In *EVA London 2012: Electronic Workshops in Computing* (2012), British Computer Society (BCS), pp. 24–33.

[71] Gephi Consortium. Gephi Quick Start. `http://www.slideshare.net/gephi/gephi-quick-start`, 2010. Accessed: 2014-10-06.

[72] Gerhardt, M., Schuster, H., and Tyson, J. J. A cellular automaton model of excitable media including curvature and dispersion. *Science 247*, 4950 (1990), 1563–1566.

[73] Gibson, O. Shopper's eye view of ads that pass us by. `http://www.theguardian.com/media/2005/nov/19/advertising.marketingandpr`, 2005. Accessed: 2016-01-19.

[74] Gino, F., and Ariely, D. The dark side of creativity: original thinkers can be more dishonest. *Journal of personality and social psychology 102*, 3 (2012), 445.

[75] Gladwell, M. *The tipping point: How little things can make a big difference.* Little, Brown, 2006.

[76] Gracia-Ibáñez, V., and Vergara, M. Applying action research in cad teaching to improve the learning experience and academic level. *International Journal of Educational Technology in Higher Education 13*, 1 (2016), 1.

[77] Granovetter, M. S. The strength of weak ties. *American journal of sociology* (1973), 1360–1380.

[78] Group, M. M. World Internet Users Statistics and 2015 World Population Stats. `http://www.internetworldstats.com/stats.htm`, 2015. Accessed: 2015-07-27.

[79] Guare, J. *Six Degrees of Separation: A Play.* Vintage Books, New York, 1990.

[80] Hansen, D., Shneiderman, B., and Smith, M. A. *Analyzing social media networks with NodeXL: Insights from a connected world.* Morgan Kaufmann, 2010.

[81] Heider, F. Attitudes and cognitive organization. *The Journal of psychology 21*, 1 (1946), 107–112.

[82] Hess, G. Disease in metapopulation models: implications for conservation. *Ecology 77*, 5 (1996), 1617–1632.

[83] Hopfield, J. J., and Herz, A. V. Rapid local synchronization of action potentials: Toward computation with coupled integrate-and-fire neurons. *Proceedings of the National Academy of Sciences 92*, 15 (1995), 6655–6662.

[84] Hopkins, B., and Wilson, R. The truth about kõnigsberg. *The College Mathematics Journal 35*, 3 (2004), 198.

[85] Houjou, K. . Development of evaluation methodology for appropriate 3d-cad practice on mechanical design education. *Journal of JSEE 61*, 2 (2013), 7–11.

[86] Hu, B., Liu, Y., and Gao, S. Parallel global optimal approach of feedback for 3d cad model retrieval. In *Computer-Aided Design and Computer Graphics, 2007 10th IEEE International Conference on* (2007), IEEE, pp. 132–137.

[87] Huang, J., Chen, Y., Zhang, Z., and Xie, Y. An affordance-integrated approach for design knowledge reuse. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* (2015), 0954406215578702.

[88] Huang, J., Piech, C., Nguyen, A., and Guibas, L. Syntactic and functional variability of a million code submissions in a machine learning mooc. In *AIED 2013 Workshops Proceedings Volume* (2013), Citeseer, p. 25.

[89] Hunt, E. Facebook brings the world three-and-a-bit degrees of separation closer. {http://www.theguardian.com/technology/2016/feb/05/facebook-brings-the-world-three-and-a-bit-degrees-of-separation-closer}, 2016. Accessed: 2016-02-08.

[90] Hyman, J. M., and Stanley, E. A. Using mathematical models to understand the aids epidemic. *Mathematical Biosciences 90*, 1 (1988), 415–473.

[91] i Cancho, R. F., and Solé, R. V. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences 268*, 1482 (2001), 2261–2265.

[92] Ip, C. Y., and Gupta, S. K. Retrieving matching cad models by using partial 3d point clouds. *Computer-Aided Design and Applications 4*, 5 (2007), 629–638.

[93] Itskovich, A., and Tal, A. Surface partial matching and application to archaeology. *Computers & Graphics 35*, 2 (2011), 334–341.

[94] Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y., and Ramani, K. Shape-based searching for product lifecycle applications. *Computer-Aided Design 37*, 13 (2005), 1435–1446.

[95] Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y., and Ramani, K. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design 37*, 5 (2005), 509–530.

[96] Jain, V., and Zhang, H. A spectral approach to shape-based retrieval of articulated 3d models. *Computer-Aided Design 39*, 5 (2007), 398–407.

[97] Jayanti, S., Kalyanaraman, Y., Iyer, N., and Ramani, K. Developing an engineering shape benchmark for cad models. *Computer-Aided Design 38*, 9 (2006), 939–953.

[98] Jeffrey, C. The students who feel they have the right to cheat. *BBC News Magazine* (2014). Accessed: 2014-11-10.

[99] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. The large-scale organization of metabolic networks. *Nature 407*, 6804 (2000), 651–654.

[100] Jones, C. Duplicate Parts…everyone has them! - Actify. `http://www.actify.com/2012/09/duplicate-parts-everyone-has-them/`, 2012. Accessed: 2013-07-26.

[101] Jun, Y., Raja, V., and Park, S. Geometric feature recognition for reverse engineering using neural networks. *The International Journal of Advanced Manufacturing Technology 17*, 6 (2001), 462–470.

[102] Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology 22*, 3 (1969), 437–467.

[103] Kazhdan, M., Chazelle, B., Dobkin, D., Funkhouser, T., and Rusinkiewicz, S. A reflective symmetry descriptor for 3d models. *Algorithmica 38*, 1 (2004), 201–225.

[104] Keegan, S. I know what the caged bird wrote: Finding the line between intertextuality and plagiarism. In *6th International Integrity and Plagiarism Conference* (2014).

[105] Kirkman, T. P. On the representation and enumeration of polyhedra. *Memoirs, Manchester Lit. Phil. Soc 12* (1855), 47–70.

[106] Knuth, D. E., Knuth, D. E., and Knuth, D. E. *The Stanford GraphBase: a platform for combinatorial computing*, vol. 37. Addison-Wesley Reading, 1993.

[107] Konig, D. *Theorie der endlichen und unendlichen Graphen.* American Mathematical Soc., 2001.

[108] Kopena, J., and Regli, W. C. Functional modeling of engineering designs for the semantic web. *IEEE Data Eng. Bull. 26*, 4 (2003), 55–61.

[109] Kretzschmar, M., and Morris, M. Measures of concurrency in networks and the spread of infectious disease. *Mathematical biosciences 133*, 2 (1996), 165–195.

[110] Krüger, D. B., and Wartzack, S. Web-based assessment of cad data in undergraduate design education. In *ASME 2014 12th Biennial Conference on Engineering Systems Design and Analysis* (2014), American Society of Mechanical Engineers, pp. V001T08A003–V001T08A003.

[111] Kuo, C.-T., and Cheng, S.-C. 3d model retrieval using principal plane analysis and dynamic programming. *Pattern Recognition 40*, 2 (2007), 742–755.

[112] Kuramoto, Y. *Chemical oscillations, waves, and turbulence*, vol. 19. Springer Science & Business Media, 2012.

[113] Lankow, J., Ritchie, J., and Crooks, R. *Infographics: The power of visual storytelling.* John Wiley & Sons, 2012.

[114] Larabi, S. Textual description of shapes. *Journal of Visual Communication and Image Representation 20*, 8 (2009), 563–584.

[115] Lee, H. Comparison between traditional and web-based interactive manuals for laboratory-based subjects. *International Journal of Mechanical Engineering Education 30*, 4 (2002), 307–314.

[116] Leng, B., and Xiong, Z. Modelseek: an effective 3d model retrieval system. *Multimedia Tools and Applications 51*, 3 (2009), 935–962.

[117] Li, L., Wang, H., Chin, T.-J., Suter, D., and Zhang, S. Retrieving 3d cad models using 2d images with optimized weights. In *Image and Signal Processing (CISP), 2010 3rd International Congress on* (2010), vol. 4, IEEE, pp. 1586–1589.

[118] Li, W., Lu, W. F., Fuh, J. Y., and Wong, Y. Collaborative computer-aided designresearch and development status. *Computer-Aided Design 37*, 9 (2005), 931–940.

[119] Li, X., and Godil, A. Investigating the bag-of-words method for 3d shape retrieval. *EURASIP Journal on Advances in Signal Processing 2010* (2010), 5.

[120] Lmaati, E. A., El Oirrak, A., Kaddioui, M. N., Ouahman, A. A., and Sadgal, M. 3d model retrieval based on 3d discrete cosine transform. *Int. Arab J. Inf. Technol. 7*, 3 (2010), 264–270.

[121] Longini, I. M. A mathematical model for predicting the geographic spread of new infectious agents. *Mathematical Biosciences 90*, 1-2 (1988), 367–383.

[122] Ma, L., Huang, Z., and Wang, Y. Automatic discovery of common design structures in cad models. *Computers & Graphics 34*, 5 (2010), 545–555.

[123] Ma, L., Huang, Z., and Wu, Q. Extracting common design patterns from a set of solid models. *Computer-Aided Design 41*, 12 (2009), 952–970.

[124] Mademlis, A., Daras, P., Tzovaras, D., and Strintzis, M. G. 3d object retrieval using the 3d shape impact descriptor. *Pattern Recognition 42*, 11 (2009), 2447–2459.

[125] Martin, I., Stubbs, M., and Troop, H. Weapons of mouse destruction: A 3d strategy for combating cut-and-paste plagiarism using the jisc plagiarism advisory service. In *2nd International Plagiarism Conference* (2006).

[126] McCandless, D. *Information is beautiful.* Collins London, UK, 2009.

[127] McCandless, D. The beauty of data visualization. In *TED Talk* (2010).

[128] McCandless, D. *Knowledge is beautiful.* William Collins, 2014.

[129] Meece, S. A bird's eye view–of a leopard's spots the çatalhöyük mapand the development of cartographic representation in prehistory. *Anatolian studies 56* (2006), 1–16.

[130] Milgram, S. The small world problem. *Psychology today 2*, 1 (1967), 60–67.

[131] Mill, F., Anderson, E., Sherlock, A., Corney, J., and Paterson, D. Network theoretic depictions and metrics for collections of 3D CAD models. Submitted, returned for review, 2013.

[132] Mill, F., and Sherlock, A. Biological analogies in manufacturing. *Computers in Industry 43*, 2 (2000), 153–160.

[133] Mitchell, M. Complex systems: Network thinking. *Artificial Intelligence 170*, 18 (2006), 1194–1212.

[134] Mol, L. The potential role for infographics in science communication. *Master's thesis, Biomedical Sciences, Vrije Universiteit, Amsterdam, Netherlands* (2011).

[135] Morales, L., and Dominguez, A. S. Assessment for learning: How plagiarism could be used as an efficient learning tool. *International Journal of Learning, Teaching and Educational Research 12*, 1 (2015).

[136] Moreno, J. Inter-personal therapy and the psychopathology of inter-personal relations. *Sociometry* (1937), 9–76.

[137] Moreno, J. L. Sociometry in relation to other social sciences. *Sociometry 1*, 1/2 (1937), 206–219.

[138] Myers, T. Faster, Better, Cheaper. `http://www.actify.com/2012/03/faster-better-cheaper/`, 2012. Accessed: 2013-07-26.

[139] Newman, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences 98*, 2 (2001), 404–409.

[140] Newman, M. E. The structure and function of complex networks. *SIAM review 45*, 2 (2003), 167–256.

[141] Paoletti, L. Leonard euler's solution to the königsberg bridge problem. *URL http://www. maa. org/press/periodicals/convergence/leonard-eulers-solution-to-the-konigsberg-bridge-problem.(Cited on pages 9 and 10)* (2011).

[142] Park, C. Rebels without a clause: Towards an institutional framework for dealing with plagiarism by students. *Journal of further and Higher Education 28*, 3 (2004), 291–306.

[143] Paulisch, F. N., and Tichy, W. F. Edge: An extendible graph editor. *Software: Practice and Experience 20*, S1 (1990), S63–S88.

[144] Philipp-Foliguet, S., Jordan, M., Najman, L., and Cousty, J. Artwork 3d model database indexing and classification. *Pattern Recognition 44*, 3 (2011), 588–597.

[145] Piegl, L. A. Ten challenges in computer-aided design. *Computer-Aided Design 37*, 4 (2005), 461–470.

[146] Playfair, W. *The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century.* T. Burton, 1801.

[147] Porter, M. Misrepresentation and visual quotations in Art and Design: A pragmatic approach. In *6th International Integrity and Plagiarism Conference* (2014).

[148] Prell, C. *Social network analysis: History, theory and methodology.* Sage, 2012.

[149] Rea, H., Sung, R., Corney, J., Clark, D., and Taylor, N. Interpreting three-dimensional shape distributions. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219*, 6 (2005), 553–566.

[150] Rea, H. J., Corney, J. R., Clark, D. E., and Taylor, N. K. A surface partitioning spectrum (sps) for retrieval and indexing of 3d cad models. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on* (2004), IEEE, pp. 167–174.

[151] Reuter, M., Wolter, F.-E., and Peinecke, N. Laplace–beltrami spectra as shape-dnaof surfaces and solids. *Computer-Aided Design 38*, 4 (2006), 342–366.

[152] Robertson, B. F., Walther, J., and Radcliffe, D. F. Creativity and the use of cad tools: Lessons for engineering design education from industry. *Journal of Mechanical Design 129*, 7 (2007), 753–760.

[153] Rodriguez, M. A. On Graph Computing. https://markorodriguez.com/2013/01/09/on-graph-computing/, 2013. Accessed: 2014-05-11.

[154] Rossignac, J. Education-driven research in cad. *Computer-Aided Design 36*, 14 (2004), 1461–1469.

[155] Rossiter, J. Which technology can really enhance learning within engineering? *International Journal of Electrical Engineering Education 48*, 3 (2011), 231–244.

[156] Ruthven, A. *Kevin Bacon is the Center of the Universe.* Google groups, 1994. Accessed: 2015-10-03.

[157] Sanna, A., Lamberti, F., Paravati, G., and Demartini, C. Automatic assessment of 3d modeling exams. *Learning Technologies, IEEE Transactions on 5*, 1 (2012), 2–10.

[158] Sant, T. Student online research and critical thinking: Wikipedia in Education. In *6th International Integrity and Plagiarism Conference* (2014), Wikipedia.

[159] Santini, S., and Jain, R. Similarity matching. In *Recent Developments in Computer Vision.* Springer, 1995, pp. 571–580.

[160] Sarkar, M., and Brown, M. H. Graphical fisheye views. *Communications of the ACM 37*, 12 (1994), 73–83.

[161] Saupe, D., and Vranić, D. V. *3D model retrieval with spherical harmonics and moments.* Springer, 2001.

[162] Scheiner, C. *Rosa Ursina sive sol.* Bracciano: Andreas Phaeus, 1626.

[163] Schepisi, F. Six Degrees of Seperation, 1993. Film, Metro-Goldwyn-Mayer.

[164] Sedghi, A. Facebook: 10 years of social networking, in numbers. http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics, 2014. Accessed: 2015-07-27.

[165] Sergey Edunov, Diuk, C., Filiz, I. O., Bhagat, S., and Burke, M. Three and a half degrees of separation. https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/, 2016. Accessed: 2016-02-08.

[166] Shilane, P., Min, P., Kazhdan, M., and Funkhouser, T. The princeton shape benchmark. In *Shape modeling applications, 2004. Proceedings* (2004), IEEE, pp. 167–178.

[167] Shlens, J. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100* (2014).

[168] Shneiderman, B., and Aris, A. Network visualization by semantic substrates. *Visualization and Computer Graphics, IEEE Transactions on 12*, 5 (2006), 733–740.

[169] Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., and Dickinson, S. Retrieving articulated 3-d models using medial surfaces. *Machine vision and applications 19*, 4 (2008), 261–275.

[170] Silventoinen, A., Denger, A., Lampela, H., and Papinniemi, J. Challenges of information reuse in customer-oriented engineering networks. *International Journal of Information Management 34*, 6 (2014), 720–732.

[171] Simon. How well do academic integrity policies and procedures apply to non-text assessments? In *6th International Integrity and Plagiarism Conference* (2014).

[172] Sivaloganathan, S., and Shahin, T. Design reuse: an overview. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 213*, 7 (1999), 641–654.

[173] Smiciklas, M. *The power of infographics: Using pictures to communicate and connect with your audiences.* Que Publishing, 2012.

[174] Smith, D. Microsoft proves there are just six degrees of separation between us. `http://www.theguardian.com/technology/2008/aug/03/internet.email`, 2008. Accessed: 2015-10-03.

[175] Smith, L. I. A tutorial on principal components analysis. *Cornell University, USA 51*, 52 (2002), 65.

[176] Smith, M., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., Dunne, C. NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010. `http://nodexl.codeplex.com/`, 2010.

[177] Srikanth, M., and Asmatulu, R. Modern cheating techniques, their adverse effects on engineering education and preventions. *International Journal of Mechanical Engineering Education 42*, 2 (2014), 129–140.

[178] Stappenbelt, B. Plagiarism in mechanical engineering education: a comparative study of international and domestic students. *International Journal of Mechanical Engineering Education 40*, 1 (2012), 24–41.

[179] Strogatz, S. H., Stewart, I., et al. Coupled oscillators and biological synchronization. *Scientific American 269*, 6 (1993), 102–109.

[180] Sylvester, J. Chemistry and algebra. *Nature 17*, 432 (1878), 284.

[181] Szykman, S., Racz, J. W., and Sriram, R. D. The representation of function in computer-based design. In *Proceedings of the 1999 ASME design engineering technical conferences (11th international conference on design theory and methodology)* (1999), Citeseer.

[182] Tangelder, J. W., and Veltkamp, R. C. A survey of content based 3d shape retrieval methods. *Multimedia tools and applications 39*, 3 (2008), 441–471.

[183] Taylor, P. The Bridges of Königsberg. What Ever Happened to Those Bridges? http://web.archive.org/web/20120319074335/http://www.amt.canberra.edu.au/koenigs.html, 2007. Accessed: 2015-07-22.

[184] Telesford, Q. K., Joyce, K. E., Hayasaka, S., Burdette, J. H., and Laurienti, P. J. The ubiquity of small-world networks. *Brain connectivity 1*, 5 (2011), 367–375.

[185] Tesson, K. J. *Dynamic Networks. An interdisciplinary study of network organization in biological and human social systems.* PhD thesis, University of Bath, 2006.

[186] TheUniversityofEdinburgh. Plagiarism. http://www.ed.ac.uk/academic-services/staff/discipline/plagiarism. Accessed: 2015-12-14.

[187] TheUniversityofEdinburgh. All the colours of nature inspire new show. http://www.ed.ac.uk/news/all-news/talbot-150312, 2012. Accessed: 2016-01-25.

[188] Tierny, J., Vandeborre, J.-P., and Daoudi, M. Partial 3d shape retrieval by reeb pattern unfolding. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 41–55.

[189] Titcomb, J. Facebook says there are only 3.57 degrees of separation. `http://www.telegraph.co.uk/technology/2016/02/04/facebook-says-there-are-actually-357-degrees-of-separation/`, 2016. Accessed: 2016-02-08.

[190] Tomiyama, T., Gu, P., Jin, Y., Lutters, D., Kind, C., and Kimura, F. Design methodologies: Industrial and educational applications. *CIRP Annals-Manufacturing Technology 58*, 2 (2009), 543–565.

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

[192] Wang, J., Cai, H., and He, Y. A new shape signature for 3d model similarity assessment. In *2008 Congress on Image and Signal Processing* (2008).

[193] Wang, Y., Liu, R., Baba, T., Uehara, Y., Masumoto, D., and Nagata, S. An images-based 3d model retrieval approach. In *Advances in Multimedia Modeling*. Springer, 2008, pp. 90–100.

[194] Watts, D. J. *Six Degrees: The Science of a Connected Age.* WW Norton & Company, 2004.

[195] Watts, D. J., and Dodds, P. S. Influentials, networks, and public opinion formation. *Journal of consumer research 34*, 4 (2007), 441–458.

[196] Watts, D. J., Dodds, P. S., and Newman, M. E. Identity and search in social networks. *science 296*, 5571 (2002), 1302–1305.

[197] Watts, D. J., and Strogatz, S. H. Collective dynamics of small-worldnetworks. *nature 393*, 6684 (1998), 440–442.

[198] Wei, L., and Yuanjun, H. Representation and retrieval of 3d cad models in parts library. *The International Journal of Advanced Manufacturing Technology 36*, 9-10 (2007), 950–958.

[199] Whitehead, J. *How do I improve my practice? Creating a discipline of education through educational enquiry.* PhD thesis, University of Bath, 1999.

[200] Wiesenfeld, K. New results on frequency-locking dynamics of disordered josephson arrays. *Physica B: Condensed Matter 222*, 4 (1996), 315–319.

[201] Winfree, A. T. *The geometry of biological time*, vol. 12. Springer Science & Business Media, 2001.

[202] Xiaoliang, B., Shusheng, Z., and Kaixing, Z. A method of 3d cad model retrieval based on genetic algorithm. In *Electronics and Information Engineering (ICEIE), 2010 International Conference On* (2010), vol. 1, IEEE, pp. V1–562.

[203] Xu, Q., Ong, S., and Nee, A. Evaluation of product performance in product family design re-use. *International Journal of Production Research 45*, 18-19 (2007), 4119–4141.

[204] Ye, X., Peng, W., Chen, Z., and Cai, Y.-Y. Today's students, tomorrow's engineers: an industrial perspective on cad education. *Computer-Aided Design 36*, 14 (2004), 1451–1460.

[205] You, C.-F., and Tsai, Y.-L. 3d solid model retrieval for engineering reuse based on local feature correspondence. *The International Journal of Advanced Manufacturing Technology 46*, 5-8 (2010), 649–661.

[206] Zhang, J., Siddiqi, K., Macrini, D., Shokoufandeh, A., and Dickinson, S. Retrieving articulated 3-d models using medial surfaces and their graph spectra. In *Energy minimization methods in computer vision and pattern recognition* (2005), Springer, pp. 285–300.

[207] Zhang, L., and Tu, W. Six degrees of separation in online society. In *Proceedings of the WebSci'09* (Athens, Greece, 2009), Society On-Line.

[208] Zhang, S.-h., Yang, C., and Thomas, S. Design knowledge and process management method based on 3d cad system. *JDIM 12*, 3 (2014), 192–200.

[209] Zheng, X.-J., Wang, Y.-S., Teng, H.-F., and Qu, F.-Z. Local scale-based 3d model retrieval for design reuse. *The International Journal of Advanced Manufacturing Technology 43*, 3-4 (2009), 294–303.

[210] Zhu, K., San Wong, Y., Loh, H. T., and Lu, W. F. 3d cad model retrieval with perturbed laplacian spectra. *Computers in industry 63*, 1 (2012), 1–11.

[211] Zhu, K., Wong, Y., Lu, W. F., and Loh, H. T. 3d cad model matching from 2d local invariant features. *Computers in industry 61*, 5 (2010), 432–439.